

大数据时代的历史机遇

产业变革与数据科学

Big Data Revolution:

How Data Science Changes the World

赵国栋 易欢欢 糜万军 鄂维南 著

清华大学出版社

大数据时代的历史机遇

——产业变革与数据科学

赵国栋　易欢欢　糜万军　鄂维南　著

清华大学出版社
北　京

内 容 简 介

大数据正以前所未有的速度，颠覆人们探索世界的方法、驱动产业间的融合与分立。本书力图系统、全面的阐述大数据在社会、经济、科学研究等方方面面的影响，或许可以帮助大家澄清一些认知误区，有助于大数据在各行各业落地生根。全书分为三大部分，第一部分重点讲述大数据时代产业发展的三大趋势以及驱动产业融合、升级、转型的根本因素，并给出践行大数据的最佳范式。第二部分首次完整阐述"数据科学"的基础性价值，论述数据科学对科学研究、社会研究、产业发展的影响，并提出数据科学的教育体系。第三部分全景式的介绍重点国家、经济体、新兴企业在大数据领域取得的进展，展示一幅真实的大数据图景，把判断留给读者，看谁拥有未来！

本书面向资本市场、产业界和学术界，成为链接三方的纽带。有助于投资人了解产业趋势、评估公司价值；有助于产业界确立公司战略方向；有助于学术界了解产业需求，促进产学的协作。

图书在版编目（CIP）数据

大数据时代的历史机遇：产业变革与数据科学 / 赵国栋等著. —北京：清华大学出版社，2013

ISBN 978-7-302-32535-2

Ⅰ. ①大…　Ⅱ. ①赵…　Ⅲ. ①数据管理-研究　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2013）第 104803 号

责任编辑：夏兆彦
封面设计：胡文航
责任校对：胡伟民
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：170mm×230mm　**印　张：**26.5　**插页：**2　**字　数：**500 千字
版　　次：2013 年 6 月第 1 版　**印　次：**2013 年 6 月第 1 次印刷
印　　数：1～15000
定　　价：49.00 元

产品编号：049306-01



赵国栋

“数据成为资产”

中国计算机学会大数据专家委员会委员、中关村大数据产业联盟秘书长、宏源证券研究所高级分析师、中国建投投资研究院特约研究员，前神州数码系统集成服务有限公司咨询总监。15 年的信息产业工作背景，在移动互联网、云计算、大数据、互联网金融等新兴领域拥有深刻、独到的见解。

邮箱：zhaogd@gmail.com

易欢欢

“没有大数据的云计算，就是房地产的代名词”



宏源证券研究所副所长、中国建投投资研究院特约研究员、前国金证券计算机行业首席分析师、前甲骨文战略咨询部高级经理、北京著名的青年财经沙龙、TMT 沙龙发起人。多次获得证券行业最高奖项新财富奖、水晶球奖金牌分析师称号。

邮箱：yisiyuan@gmail.com



糜万军

“数据之和的价值，
远远大于数据价值之和”

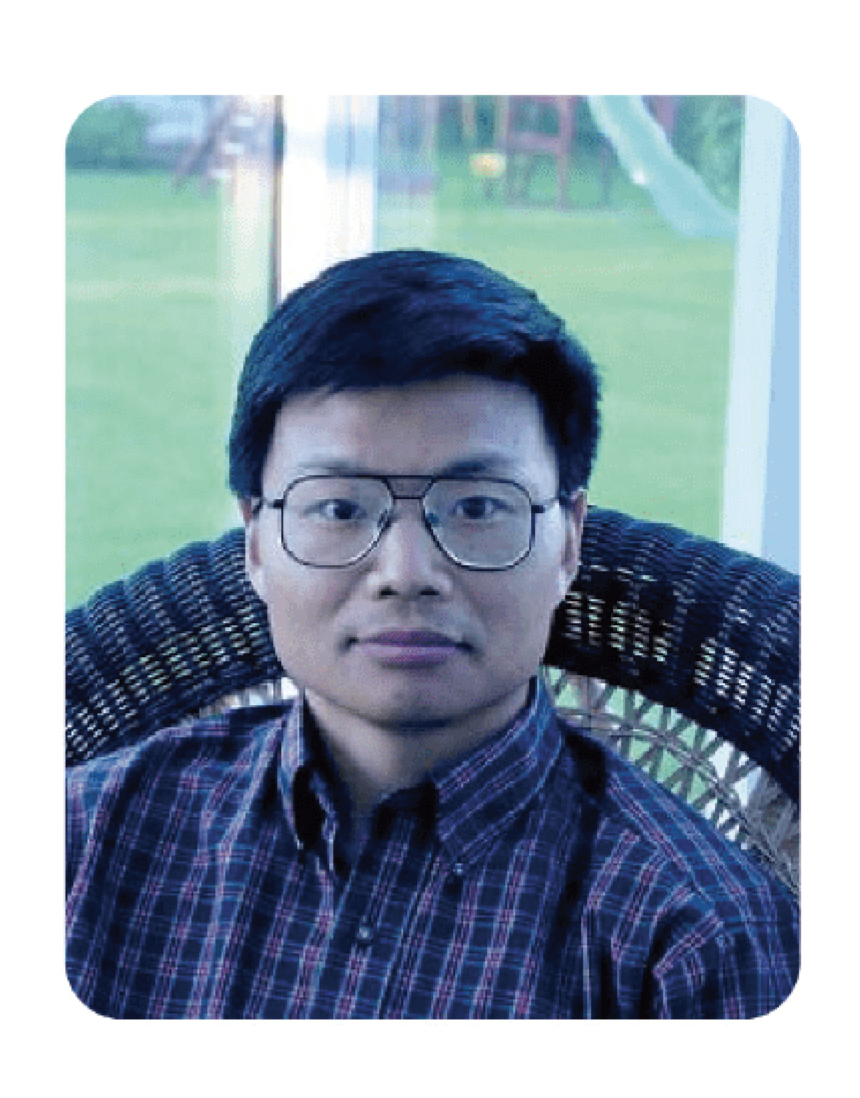
现正在创建大数据技术公司。研究方向主要包括高性能计算和大规模数据挖掘。荣获“2011 中关村高端领军人才”、“2012 中关村十大海归新星”等称号。

邮箱：wanjunmi@gmail.com

鄂维南

“数据科学将达到与自然科学分庭抗礼的地位”

中国科学院院士
北京大学长江讲座教授
美国普林斯顿大学教授





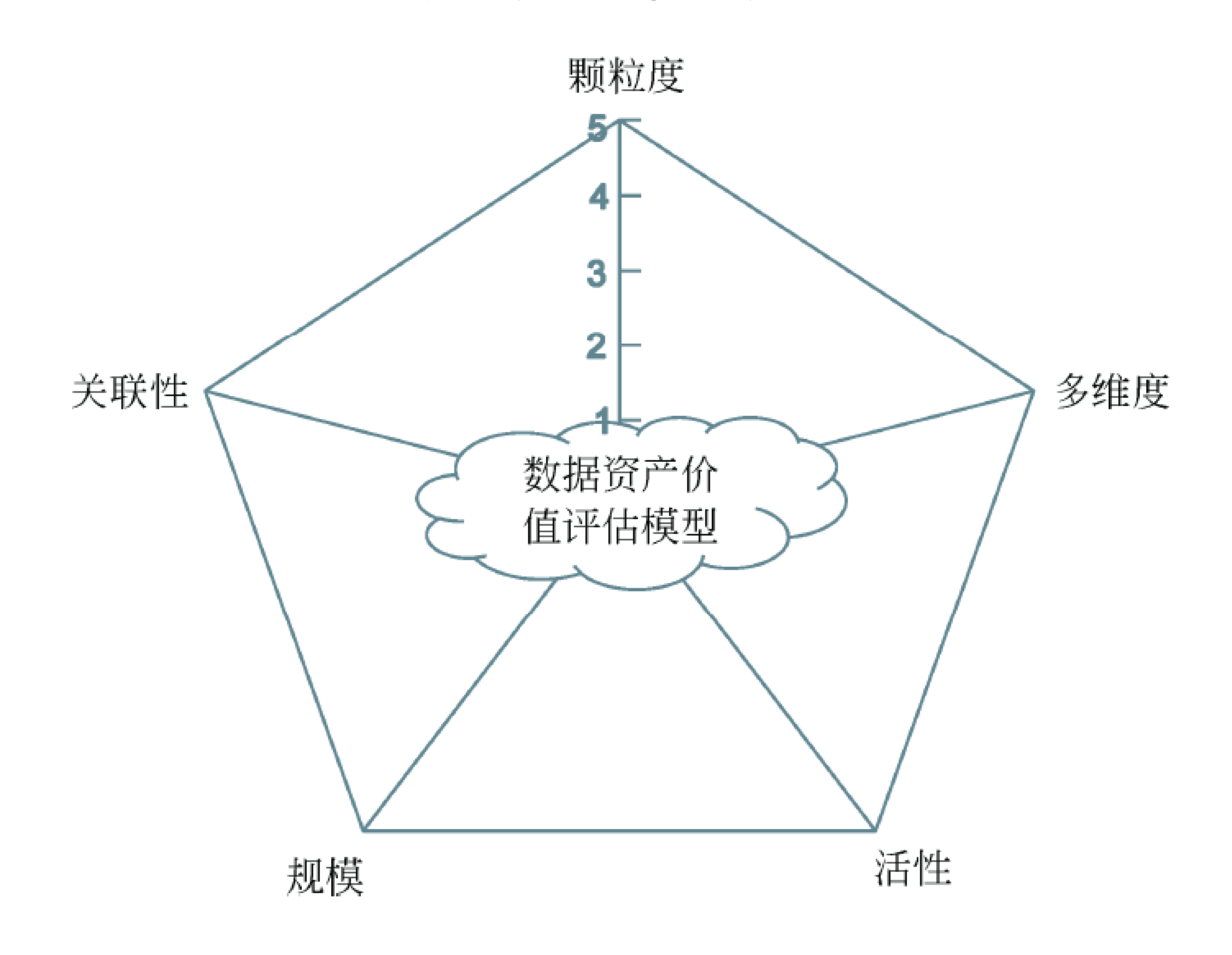




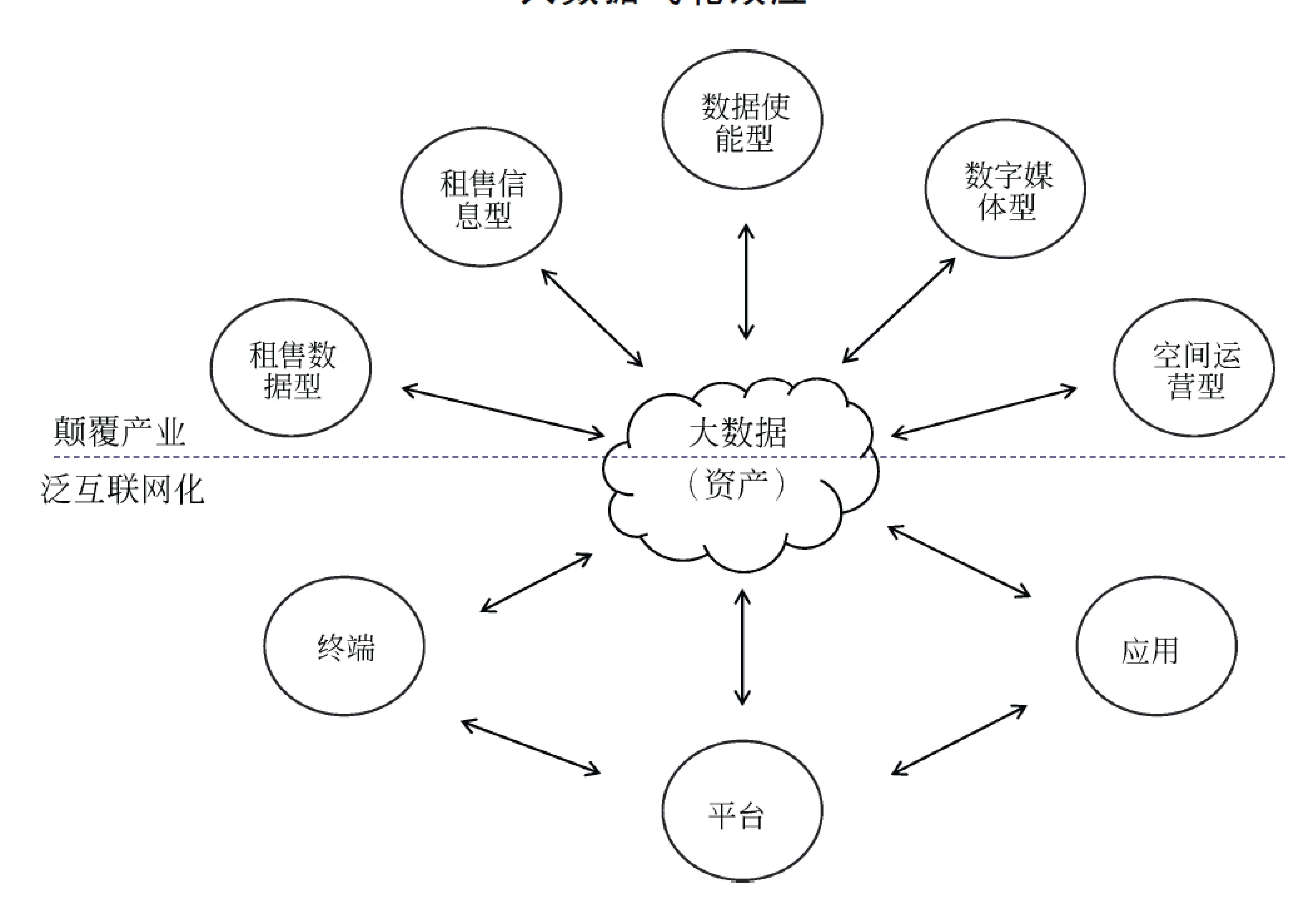


大数据的认知框架
智慧的商业
数据成为资产
公司价值与其拥有的数据规模、活性，以及其解释、运用数据的能力成正比
数据资产演绎精彩的商业模式，推动公司变革
泛互联网化
大数据
行业垂直整合
越靠近最终用户和消费者的公司，在产业链上拥有越来越大的发言权
网络功能内生为应用的API
终端的价值和其数据流量与活性成正比
海量数据
类型繁多
用途多样
稀缺珍贵
快速获得
Hadoop
大融合数据
TB+
PB+
大交易数据
大行为数据
Oracle
DB2
Teradata
移动互联
社会化网络
车联网
GIS
云计算
物联网

数据资产评估模型



大数据飞轮效应



序一 FIRST

2012 年，我个人的认识水平有一次重要的提高，那就是关于大数据的了解和认识。这个了解和认识的发蒙，来自于本书的两个作者赵国栋和易欢欢。在我的印象中，中国资本市场上最先发出“大数据”声音的，就是他们两个人。2012 年初，和君请他俩来做一个大数据主题讲座，一场讲座下来，瞬即为和君的咨询师们打开了一扇新的窗户，我们开始以大数据的眼光，重新看待企业战略、产业整合、商业模式、竞争要害、企业关键成功要素、核心能力、企业组织变革、企业与客户的关系等等重要概念或命题。这些概念和命题，在传统管理学里，都有清晰的界定和意义。赵国栋、易欢欢展开的大数据时代画卷，让我们意识到一个紧迫而严重问题：如果一个企业或一个管理咨询公司的知识和概念，还停留在传统管理学对这些命题的理解中，我们将彻底 Out。跟上时代的步伐，及时完成知识和理念上的更新，没有什么比这更重要的了。从此，和君咨询的公司发展取向、核心能力构筑、咨询师培训的课程设计、和君为客户提供的咨询建议、和君商学的教学安排，开始有了大数据的思维。我深知，这对和君公司、对和君员工、对和君客户、对和君学子，都有着战略性的意义。我觉得赵国栋和易欢欢，是和君的贵人。

经我推荐，恒安国际董事会邀请易欢欢和赵国栋专程飞赴香港为恒安全体董事作了一次“大数据与传统产业升级”的报告。我在董事会上聆听了全程，最大的感触是，各行各业，尤其是传统产业都面临着在大数据和移动互联网时代如何彻底转型和再造问题。我喊了十几年的产业整合，也在大数据时代出现了全新的整合逻辑和实现契机。正如这本书的宣传定位语所言：缺少数据资源，无以谈产业；缺少数据思维，无以言未来。恒安国际总裁许连捷先生听完赵国栋、易欢欢的报告后，就大数据思维对恒安、对快速消费品行业意味着什么，作了敏锐、深刻、快速反应的评论和强调。恒安主营卫生巾、纸巾、纸尿裤、休闲食品，地道的传统产业，年届花甲的总裁，对大数据思想的敏感和快速反应，令我印象十分深刻。我仿佛看到了一个未来景象：各行各业的传统产业，都可能在大数据和移动互联时代，重现生机、焕发青春。当然，与此对应的是，凡是不能跟上时代步伐的企业和行业，命运就是永久地走进过去，退出未来的舞台。

赵国栋、易欢欢，都是典型的理工男。在认识我之前，他们分别在神州数码和Oracle从事理工男的工作。2008年他俩考入我办的和君商学院，开始接触商学，关注金融，自此看到了“理工男”之外的全新的商业世界。从和君商学院毕业后，他们开始进入证券行业，从事IT行业的分析师岗位。他们理工知识基础好，作风踏实，十分敬业和勤奋，很快就脱颖而出了，多次获得证券分析师行业里的重要奖项，比如新财富行业分析师排名第一、水晶球奖金牌分析师等称号，更重要的是他们的思想认识和专业水平，提高很快、进步很大。我作为他们曾经的老师、作为他们从理工男转入证券行业的引路人，感到很欣慰、很赞赏。真的没想到，短短的3年时间他们就可以用思想认识和专业水平来反哺和君、提升老师了。现在，在大数据问题上，我是他们的学生，还需要向他们多多学习、持续学习。这本书，堪作我学习的课本，一个贯通技术理解、产业认识和资本市场估值的难得教材。而技术大牛糜万军先生和中国科学院院士鄂维南教授亲自参与这本书的写作，更让我觉得弥足珍贵。据说本书的“数据科学”章节，是鄂维南院士用纸和笔，一字一句地写出来的，然

后再由工作人员敲打成电子版本。这一细节，让我对鄂院士的认真感到肃然起敬，也掀起了我对笔墨写书时代的一种怀旧，俨然像对某种古典而失传工艺的隐隐恋想。

兹为序。

王明夫博士/和君咨询董事长

2013 年初夏，于北京和君咨询

序二 SECOND

精准、全面、及时和“数字会说话”一直是人们对企业和政府等公共组织信息系统处理信息的愿望。在之前的信息技术和产业模式条件下我们实现了这些愿望的一部分，但远没有达到人们对信息的理想期求。尽管我们已经提出并运用了对应的普适计算、泛在计算、实时系统和商业智能的理念和模式，问题仍然没有得到根本的解决，直至“大数据”的思想、模式、技术和产业开始真正地形成。

未来企业都将会是“数据驱动的企业”，无论你处于什么行业，企业规模大小。一些企业已经先行一步并在行业中获得巨大的领先优势，一些企业刚刚开始行动，更多的企业还在认识甚至还没有认识到的阶段。

“数据”作为企业和公共组织越来越重要的资产，就像当年“知识产权”对于企业资产形态的突破以及由此带来的企业进步发展一样，将历史性地改变着企业资产的理念和进步发展进程。

我十分赞同作者对大数据的观点，大数据不仅仅是一项技术，更是思维方式、发展战略和商业模式。

作者洞察的“行业垂直整合”趋势，以及提出的“终端”+“应用”+“平台”

以及“数据”四位一体的泛互联网化范式对我们产业企业、产业主管部门以及投资机构都具有重要的指引意义。

正是在数据驱动企业发展、驱动社会发展的全新时代背景下，“数据科学”显得尤为重要和迫切，不仅仅对于学术研究界，更对于企业和社会实务界，同时也需要产学研进一步的深度合作。

“大数据”是最近两年来的一个热词，市面关于大数据的书已经有很多，但是从产业的角度展开并深入剖析的，这是我看到的第一本，这本书让大数据不再停留在理念和技术的层面，而是深入到商业价值与模式、产业机构与形态的层面，读着更具现实感。

本书呈现的特色，与作者的背景和结构不无关系。赵国栋和易欢欢先生是国内最优秀的计算机行业分析师之一，他们在 2012 年初，率先把大数据概念介绍到中国资本市场，他们让我们看到行业分析师洞察产业的独特视角和优势；糜万军先生是我们企业界的顶尖技术专家，是中关村的高端领军人才；鄂维南先生是中国科学院院士，是我们尊敬的中国数据科学领军人。这是资本市场、产业、学术三个方面的专家通力合作的第一本大数据方面的书，书的专业素养自然经得起挑剔，不乏智慧的火花；在可读性方面，几位作者也是用心架构，深入浅出，案例丰富。

除了企业界的人士，我建议政府工作人员也应该读读这本书，因为大数据必将深刻改变政府的行为，改变政府和社会的关系；学者应该读读这本书，因为资本和产业界的人士会从不一样的角度解读大数据，给你一个全新的视角；学生应该读读这本书，因为大数据将改变企业对人才的需求，应该早做准备。

王文京/用友软件股份有限公司董事长

序三 THIRD

近年来，互联网与传统产业融合进程加速推进，传统产业的运营模式和游戏规则正在被逐步瓦解并再造。苹果、三星颠覆了传统手机终端，亚马逊、阿里巴巴、京东商城改变了传统零售业，Twitter、Facebook 和微信撼动了传统媒体社交……这样的故事正在不断上演。

信息和信息技术是金融业的关键要素，每一次创新和突破，都会给金融业带来重大的影响，这种影响甚至会比其他产业更加明显。因此，金融业基于新兴技术的升级变革已成为大势所趋。特别是云计算、物联网、社交网络、移动互联网、大数据等新兴技术层出不穷，改变了信息的生产、传播、加工和组织方式，打破了传统的信息不对称和物理区域壁垒，对金融业的生存环境和方式造成了明显的影响。具体到证券业而言，可以从两个事件中感受这种变化，一是 2012 年 5 月，社交媒体监测平台 DataSift 通过监测 Twitter 上的情感倾向准确预测了 Facebook 上市当天股价的走势；二是 2013 年 3 月中国证监会发布的《证券账户非现场开户实施暂行办法》，允许见证开户和网上开户，这对证券公司传统经纪业务以及研究业务、资产管理业务等未来发展都有长远而深刻的影响。

当然，这种影响并不仅局限于证券业，整个金融业无一例外都已经或多或少感受到了这种新变化。比尔•盖茨曾经预言，“传统商业银行将成为 21 世纪最后的恐龙”，而如今商业银行并没有灭绝，且发展得还很好。但不可否认的是，借助互联网、大数据的崛起，一批新兴力量已经对银行等传统金融机构产生了一定的冲击。也许在业务量上还没有构成实质性的威胁，但这些力量所代表的新技术、新模式、新思想却不可小觑，值得所有传统金融机构去深入探究和学习。

《大数据时代的历史机遇》一书，从大数据这一视角切入，全面呈现了第三方支付、供应链金融、网络小额贷款和 P2P 网络借贷等多种新生金融业态，并深刻揭示了大数据成为继土地、人力、技术、资本之后的新型资产，是金融业未来打造核心竞争力的关键要素，对于我们思考未来金融业的发展趋势和格局构成都会有很好的启迪意义。

本书的两位主要作者——赵国栋和易欢欢，都是我非常优秀的同事，他们在相关产业前瞻性研究方面做了大量的工作，取得了不错的成绩。这次他们不仅率先将大数据引入资本市场，而且还进一步针对大数据与金融、媒体等传统产业的融合趋势进行了深入研究，并提出了很多非常具有价值的观点和意见。《大数据时代的历史机遇》可以说是他们最新研究工作的智慧结晶，是一本难得的大数据相关作品！

胡强/宏源证券总经理

前言 FOREWORD

星罗密布的人造卫星和数以千万计的各种传感器，源源不断地侦测、创建和传输大量的数据。人们的喜怒哀乐、吃穿住行等人性化的表征和行为都在虚拟的网络空间中再现和升华。人类全面进入了数据时代。数据的影响已经渗入到了产业、科研、教育、家庭和社会等各个层面。 可以说，缺乏数据资源，无以谈产业；缺乏数据思维，无以言未来。

尽管大数据已经成了一个热点话题，但目前大数据方面的文献大多聚焦在它的数据容量，数据多样性以及访问速度上，也就是所谓的三个“V”。本书则穿透数据爆炸的表象，聚焦于探讨大数据对于产业变革、科学研究的巨大影响。大数据正以前所未有的速度，颠覆人们探索世界的方法，驱动产业间的融合与分立。因而当务之急是，怎么认知大数据？如何让大数据更好地应用到科学研究中去？如何让大数据切实帮助公司突破增长的瓶颈？本书力图系统、全面地阐述大数据社会、经济、科学研究等方方面面的影响，或许可以帮助大家澄清一些认知误区，有助于大数据在各行各业落地生根。

本书分为三大部分：第一部分阐述大数据时代产业趋势的问题；第二部分重点

在于数据科学；第三部分概览世界主要国家、经济体在大数据方面的政策和举措，海外巨头以及新兴公司在大数据领域的实践。

“数据成为资产”是最核心的产业趋势。正如本书概述所提到的：“当写完这些案例，回头审视产业的起起伏伏，发现产业兴衰的决定性因素，已经不是一城一池的争夺。土地、人力、技术、资本这些传统的生产要素，甚至需要追随“数据资产”，重新进行优化配置。”那些拥有优质数据资产的公司，挟天子以令诸侯，不断地攻伐、侵袭其他产业的传统领地。产业融合大幕随之拉开，天平却向这些新兴的公司倾斜。由此笔者也得出第一个公司价值的判断标准：“大数据时代公司的价值，与其数字资产的规模、活性成正比，与其解释、运用数据的能力成正比。”

本书第一部分用四章的篇幅来描述“数据资产”，提出数据资产的评估模型，并以此为基础来判断符合哪些条件才是优质的数据资产，才具备产业跨界攻伐的潜力。围绕数据资产的运用，衍生出不同的商业模式，通过大量的学术研讨和商业案例，来阐释这些商业模式的合理性、颠覆性。第四章和第五章分别描述了已经被颠覆的媒体行业和正在受到冲击的金融行业。

具体到信息产业内部，当下另一个重要的趋势是“行业垂直整合”。那些越是靠近产业链末端，越是靠近最终消费者的公司，将在产业链中拥有越来越大的发言权。这一趋势对中国信息产业而言，意义尤其重大：它是大数据时代，我国信息产业实现弯道超车的契机。影响这个趋势的关键因素包括开源软件的兴盛、软硬一体化重新唱主角、应用为王、极简主义盛行等。洞悉行业垂直整合趋势，将对一、二级市场的投资判断，有重要的参考意义。本书第六章将重点谈论这部分内容。

泛互联网化是笔者提出的另一个主要思想，也是收集数据资产、发挥大数据商业价值的最佳实践。多种形态的设备、软件都会具备联网的功能，联网成为泛化的功能存在于各种设备、各种软件之中。笔者系统地考察了苹果、谷歌等引领世界潮流的公司商业模式，也遍访国内传统的 IT 公司，提出“终端”+“应用”+“平台”以及“数据”四位一体的泛互联网化范式，重点揭示该范式的特征与实践，批判“工

业时代的标准化思维”。灵活利用泛互联范式，传统企业会取得意料之外的高速增长，也是创业型公司从零开始积累数据资产的正途。这个话题的初步探讨参见第七章。

本书第二部分围绕“数据科学”展开。大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑战。它不仅将给现有的科研和教学体制带来大幅度的变革，也会给科学与产业之间的关系、科学与社会之间的关系带来大幅度的变革。信息时代，万物数化。许多学科已经和信息科技深度融合，形成新的研究领域，譬如生物信息学、天体信息学、数字地球、计算社会学等。“用数据来研究科学”已经是科学研究的主要手段之一。另一方面，大量的、非结构化的数据，同样需要科学的手段，来去芜存菁，即“科学的研究数据”。另外，产业界在生产经营中积累丰富的数据，学术界则有待于实践检验的模型和算法。“数据科学”为学术界和产业界的紧密衔接提供了纽带和桥梁，成为促进产、学、研深度融合的重要契机。

本书前两部分偏重构建大数据相关理论和趋势，第三部分则全景扫描各政府、各大经济体、各行业领头羊和典型的新兴公司在大数据方面的具体实践。如果没有第三部分，前两部分就像自说自话，成了无源之水。在各国政府的大数据行动中，美国的动向无疑最值得关注。第十一章几乎通篇都在讲美国政府的开放策略。大家从中可以看到，美国政府是如何利用数据技术来促使政府变得更加透明、廉洁和高效。读罢这一章，大家也会很容易理解奥巴马政府《大数据研究与发展计划》的初衷。第十二章阐述了大型公司如何利用大数据技术相互攻伐，第十三章则重点放在有哪些值得关注的新兴企业，对于专注于早期投资的机构而言，这章具备十分重要的参考意义。

这本书是笔者和易欢欢、糜万军、鄂维南院士通力合作的结晶。易欢欢先生是宏源证券研究所副所长，曾荣获 2011 年新财富奖、水晶球奖金牌分析师第一名，在资本市场首提大数据概念，引领一时之风潮。糜万军先生现正在创建大数据技术公司，同时也是“中关村高端领军人才”的代表人物。糜总在数据统计、定向广告的核心算法方面造诣深厚。鄂维南先生是中国科学院院士，同时也是北京大学长江讲

座教授、美国普林斯顿大学教授，他已倡导数据科学多年，是我国发展数据科学的领军人物之一。

本书系统地总结了笔者多年的工作心得、行业感悟。本书思想来自于产业界、学术界、政府人士的反复沟通和碰撞，成书之际，谨在此表示深深地感谢。他们是（排名不分先后）国金证券研究所副所长李伟奇、甲骨文产品战略部总监刘松、用友集团董事长王文京、拓尔思总裁施水才、启明星辰首席战略官潘柱廷、上海证券交易所总工程师白硕、神州数码 CTO 谢耘、神州数码徐拥军、民生证券 CIO 颜阳、SAP 全球数据库解决方案亚太及日本区技术总监卢东明、百度公司多媒体部副总监余凯、京东商城副总裁李曦、北京大学教授姚远、工信部电子科学技术情报研究所陈新河、工信部软件与集成电路促进中心陈越等。

感谢网友@尹锴_ink、@夏明武，他们慷慨无私地提供了大量的资料和职业感悟。感谢笔者的写作团队，他们利用业余时间收集、翻译、整理资料，校对文字。其中刘丰（第八、十一章）、闻学臣（第五章）、李隽钦（第四章、第十三章）甚至参与撰写了部分章节。笔者的写作思路和风格时常调整，导致大家许多工作成为无用功，收集大量资料却无一采用。尽管如此他们依然任劳任怨，志愿付出。他们是蒋传臣、靳松、陆安、刘丰、许文星、闻学臣、魏增、金慈航、尹佳、丁新、安征、王萌、曹宇峰、孙思远、徐湘童、王宁、吕殷楠、宋航、胡博、杨宣华、王东莹、何全、王宁、魏芳、曾奕恺、胡韦力、扈培培、赵晖、刘翔、刘笑逸、李隽钦、冯达、葛婧瑜、张中峰、张娟。

感谢摩宝时代为本书提供的二维码支持。

感谢清华大学出版社的信任与等待。

再次感谢！

作者

2013 年 1 月于北京



第一章　大数据概述

大数据是“在多样的或者大量的数据中快速获取信息的能力”，其关乎国计民生、产业兴衰、公司存亡，不可不察。

第一部分　产业大势

第二章　大数据时代已经到来

资本市场、产业界、学术界、政府都在紧锣密鼓地行动，四方联手推动 2012 年成为大数据时代的元年。

第三章　数据成为资产

大数据时代公司的价值与其数据资产的规模、活性成正比；与其解释、运用数据的能力成正比。

第四章　大数据颠覆媒体行业

传统平面媒体业正在经历历史上最严重的倒闭浪潮，取而代之的是新兴的互联网媒体公司。以谷歌为代表，他们以数据资产为中心，创造了迄今为止最完美的商业模式之一。

第五章　互联网金融

比尔·盖茨曾说:“传统银行若不能对电子化作出改变，将成为 21 世纪行将灭绝的恐龙”，从小微信贷、众筹、互联网金融等新兴的金融服务模式来看，金融业不得不经历痛苦的嬗变过程。

第六章　大数据加剧产业的垂直整合趋势

大数据时代，消费者真正登上了舞台中央。哪些越靠近最终消费者或者用户的公司，在产业链上就拥有越来越大的发言权。产业生态将围绕消费者重构。

第七章　泛互联网化是发挥大数据价值的最佳范式

那些仅仅拥有产品，无法形成终端、平台、应用、数据一体化的公司，将难逃被颠覆的命运。泛互联范式成为累积数据资产、发挥数据资产价值的最佳范式，也是构成大数据思维的重要组成部分。

第八章　大数据掀起的企业组织变革

大数据首先是一种思维方式，必须融入到企业的每一个毛细血管中。运用大数据思维必将审视企业与客户的关系，企业的战略、组织、文化都将因大数据而彻底改变。

第二部分　数据科学

第九章　数据科学

大数据在科学领域的表现是数据科学的兴起，数据科学将逐渐达到与其他自然科学分庭抗礼的地位。用数据研究科学，科学的研究数据。

第十章　数据技术：当前进展及关键问题

欲工其事必先利其器。促进大数据在各行各业落地的重要因素，除了建立大数据思维以外，必须掌握新兴的处理技术。需要重新审视企业的软件开源策略、数据处理技术、人才培育计划。

第三部分　全景扫描

第十一章　国家选择

开放、共享是大数据时代的核心精神。但是于政府而言，大数据是把双刃剑，它既能促进政府开放、透明，又能帮助加强集中管控。选择考验智慧！

第十二章　巨头碰撞

新兴的产业巨头凭借独一无二的数据资产，正在重新定义产业生态和竞争格局，老牌科技公司沦为看客，围观的传统产业逐一被颠覆。

第十三章　创新凶猛

新兴的大数据公司如雨后春笋，观察他们的成长，我们才深深体会到产业的脉动、变化的节奏和演变的方向。毫无疑问，他们正在重新定义未来。

附录　大数据发展大事记

后记

参考文献

引子

大数据总统奥巴马

2012 年 8 月份，美国总统大选正如火如荼。出人意料的是，奥巴马总统的数据团队要求他去一家叫 Reddit 的新闻网站去回答问题。对许多人来讲，Reddit 是一个陌生的名字，总统的高级助手们对它也不甚了解。但是来自数据团队的回答却非常简单：“因为我们需要动员的一些人，经常在 Reddit 上。”

这仅仅是选战过程中一件毫不起眼的数据决策案例。事实上，奥巴马的数据团队非常神秘、低调，但其触角又无处不在，几乎左右了整个大选，他们被内部人士戏称为“核编码”。他们创建了单一的巨大系统，可以将从民调专家、筹款人、选战一线员工、消费者数据库、以及“摇摆州”民主党主要选民档案的社会化媒体联系人与手机联系人那里得到的所有数据都聚合到一块。这个组合起来的巨大数据库令奥巴马的数据团队工作极富成效，令人惊叹[①]。在这个组合的数据库中，每个选民甚至被精确地划分为 1000 多个特点，通过建模和算法分析，系统能为每个选民找出

① 英文原文参见 CNN 网站 http://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team。

一个最能说服他的理由；每晚进行 6.6 万次模拟选举，在个体水平上，计算出奥巴马在任何一个摇摆州的胜率。事实上不仅如此：

他们建立的模型能够预测谁会在线捐款。

他们用来网上筹款的邮件，也充分利用了数据收集和分析。

他们借助模型帮助奥巴马筹集到创纪录的 10 亿美元。

他们帮助优化电视精准投放广告的模式。

他们创造出了摇摆州选民的精细模型。

他们计算出第一夫人发的拉票邮件在春天最受欢迎。

他们利用数据来详细分析关键州的选民。深入分析各个族群的选民在任何时刻的趋势。在总统候选人的第一次辩论之后，他们分析出哪些选民倒戈，哪些没有。

他们利用熟人效应，开发 Facebook App 拉票。

他们为竞选团队购买广告提供决策参考。

他们通过一些复杂的模型来精准定位不同选民，他们购买了一些冷门节目的广告时段，而没有采用在本地新闻时段购买广告的传统做法。广告效率相比 2008 年提高了 14%。

他们导致经验主义的竞选专家的作用急剧下降，能够分析大数据的量化分析专家和程序员的地位却大幅提升。

他们让政客们，尤其是对手知道政治领域的大数据时代已经到来。

一瓶茅台酒的旅程

消费者最头疼的恐怕还不是茅台酒的价格，而是能否买到货真价实的茅台。“道高一尺魔高一丈”，茅台历来的防假手段，除了推高茅台酒瓶的回收价格以外，似乎并没有真正让消费者放心。

为每一瓶茅台建立“档案”，消费者可以轻松方便地查询到任何一瓶茅台酒的档案材料，是防假的终极解决之道。每一瓶酒都有一个独立的“身份证号”，铭刻到酒

瓶上，在信息系统中记录下从灌装到出厂、运输、批发、零售所有环节的信息。人们只要把“身份证号”传输到网站一查，真伪立辨。这个办法看起来容易，但是真正实施，我们立刻会被淹没在大量的数据之中。

不仅仅是茅台，中国目前所有食品面临“安全、卫生”的大难题。如果能把茅台酒的做法推而广之，无疑是全民之福。但是这些海量的数据记录，对传统的信息处理技术提出了巨大的挑战。

茅台的故事，其实可以引发管理理念的变化。这是管理日益精细化的具体体现。原来“茅台们”的管理都是按照生产批次，通常认为同一个生产批次的产品，是没有差别的。现在的管理理念则不同，要求对每一件单品实行差别化管理。

城市治理中，也在发生同样的事情。小到每一个下水道井盖都被仔细编号、追踪。这当然另我们的生活更加便利，但产业界首先需要应对的则是大数据的挑战。

导读：

1. 大数据正以前所未有的速度颠覆人们探索世界的方法，引起社会、经济、学术、科研、国防、军事等领域的深刻变革。
2. 数据成为资产、产业垂直整合、泛互联网化是大数据时代的三大发展趋势。数据资产成为和土地、资本、人力并驾齐驱的关键生产要素。围绕数据资产可以演绎跌宕起伏的产业大戏。
3. 数据科学应运而生并将成为科研体系中的重要组成部分，逐渐达到与自然科学分庭抗礼的地位。数据科学既可以推动数学、计算机科学、统计学、天体信息学、生物信息学、计算社会学等学科的发展，又能够助力产业界升级转型。
4. 需要在宏观尺度拓宽大数据视野、建立完整的大数据思维；正视普遍存在的三大数据治理问题（数据割据、数据孤岛和数据质量）及人才短缺的现状。

第一章

大数据概述

大数据是“在多样的或者大量的数据中快速获取信息的能力”。

——笔者

大数据，事关国计民生、产业兴衰、公司存亡，不可不察。信息科技经过 60 余年的发展，数据（信息）已经渗透到国家治理、国民经济运行的方方面面。经济活动中很大一部分都与数据的创造、传输和使用有关。2012 年 3 月，奥巴马公布了美国《大数据研究和发展计划》[①]，标志着大数据已经成为国家战略，上升为国家意志。

国家竞争力将部分体现为一国拥有数据的规模、活性，以及解释、运用数据的能力。国家数字主权[②]体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间[③]。没有数据安全，也就没有国家安全。华为、中兴开拓美国市场受挫，就是非常明显和清晰的信号。美国政府对自家数据安全的重视程度，已经到了不能让任何外国信息基础设施产品供应商染指的地步。华为此前一直希望通过竞标和并购等方式进入北美市场，多年来未能如愿。2008 年，华为与贝恩资本联合竞购 3COM 公司，却因美国政府阻挠未能成行；2011 年，华为被迫接受美国外国投资委员会的建议，撤消收购 3Leaf 公司特殊资产的申请；同样是在 2011 年，美国商务部阻止华为参与国家应急网络项目招标。

再看美国国防部立项的几个大数据项目[④]：多尺度异常检测（ADAMS）项目，解决大规模数据集的异常检测和特征识别的问题；网络内部威胁（CINDER）计划，旨在开发新的方法来检测军事计算机网络与网络间谍活动，提高对网络威胁检测的准确性和速度；Insight 计划，主要解决目前情报、监视和侦察系统的不足，进行网络威胁的自动识别和非常规的战争行为……参见附录四。其他部门包括国土安全部、

---

① 《大数据研究和发展计划》原文网址：http://www.whitehouse.gov/blog/2012/03/29/ big-data-big-deal，中文译稿参见本书附录四。

② 通过搜索引擎，并未发现其他文献强调“数字主权”。之所以采用“数字主权”，而非“数据主权”，主要因为构成信息科技的基础是“0”、“1”两个二进制的数字。所有的数据在本质上都是“0”、“1”的排列组合。

③ 参见国金证券大数据系列报告第三篇《以数据资产为核心的商业模式》，第 1 页。

④ 原文参见 http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data _fact_sheet_final_1.pdf。

能源部、卫生和人类服务部、国家航天总局、美国国家科学基金会、美国国家安全局、美国地质调查局纷纷推出大数据项目。奥巴马指出：“通过提高我们从大型复杂的数据集中提取知识和观点的能力，加快科学与工程前进步伐，改变教学研究，加强国家安全。”

产业层面，大数据技术虽然发源于信息科技，但其影响已经远远超出信息行业。数据已经存在于全球经济中的每一个部门，就如固定资产和人力资本等生产要素一样，如果没有它许多现代经济活动根本就不会发生。笔者观察到一些新兴的互联网公司，利用新技术大规模地收集数据，预判客户行为，然后在不同的行业纵横捭阖。它们剑锋所指，现代服务业无不受其锋芒所迫，或随波逐流，或奋起反击。但缺少数据资产、缺少强大的数据分析能力，这类公司无疑处在被颠覆的边缘。笔者也看到传统行业的公司，数十年如一日坚持积累当时被视作“废料”的数据，现在回头审视这些数字化的资产，居然一跃成为人类的宝库。凭借独一无二的“数据资产[①]”，公司进入相关行业，易如反掌。

当笔者回头审视产业的起起伏伏时，就会发现决定产业兴衰的根本性因素已经不是一城一地的争夺了。土地、人力、技术、资本这些传统的生产要素，甚至需要追随“数据资产”重新进行优化配置。封建时代，往往是裂土封王，权贵都是大地主；工业革命后，制造业巨子成为偶像；资本市场化后，受到追捧的是拥有大量钱财的投资家。但是在大数据时代，“数据资产”成为最重要的生产要素，拥有大量数据资产的人，已经成为美国总统的座上宾[②]。

产业的分分合合，一直是资本市场非常喜欢的故事。不管是分拆也好，整合也罢，资本市场都有钱赚。以往产业的整合基本围绕产业链展开，要么向上游扩展，要么向下游兼并。但是在大数据时代，人们看到的商业图景是围绕“数据资产”拉

① 数据成为资产，参见国金证券大数据系列研究报告《大数据时代的三大发展趋势及投资方向》。

② 美国总统奥巴马于 2011 年 2 月 17 日与多名科技界领袖共进晚餐。总统左侧是苹果公司创始人史蒂夫·乔布斯，右侧是 Facebook 的创始人马克·卡克伯格。

开产业并购的大幕。谷歌所有的收购或者推出的新产品，都是为了增加数据资产的“维度”和“活性”[①]。所有观察公司发展、产业未来的机构或者个人，如果忽略“数据资产”，或者对“数据资产”认知肤浅，必将导致错误的判断。大数据将是决定产业未来的战略性资产。未来产业间的整合并购，将会在很大程度上围绕“数据资产”展开争夺。

企业家、投资人、咨询顾问、分析师，必须要从战略层面思考大数据对产业、公司的影响。2012 年初，笔者曾经和恒安国际的董事会一道交流大数据对制造业的影响。会上许连捷[②]总裁说：“在大数据时代我们收集数据，研究消费者行为，推出新的产品，改善供应链，降低库存。一句话就是把大数据融入到经营中去。也许有可能把库存降到近乎‘0’的水平。”所以，我们谈大数据，首先是思维方式的问题，要建立全面、系统的大数据意识，其次才是落实到公司战略。大数据对公司的影响是多方面的，涉及组织、文化、流程、技术等。本书第八章将专门详细论述大数据对公司组织结构的影响，在此不赘言。

具体到中国信息产业，发展速度一直落后于国外的巨头，长期处在产业链的末端，赚取刀片一样的利润，积累到最后发觉只形成了简单可替代的“中国制造”而非具备革命性创新性的“中国智造”。国家拿出大笔资金扶持上游环节的拓荒者，如CPU、操作系统、办公软件，但是相关领域国内外的差距过于遥远，也缺少大规模的商用市场，花了国家的钱，却鲜有在商业上大获成功的先例。但是在新兴的大数据处理领域，中外公司几乎站在同一起跑线上。中国作为数据的巨大产生国，有着更广阔的应用空间。比如，中国移动、工商银行、淘宝，已经具备世界级的产业应用环境。有业内人士表示，单纯考虑狭义的大数据处理技术（如 Hadoop、MapReduce、模式识别、机器学习等），中外差距仅有 5 年左右。如果考虑数字资产规模以及利用的技术，中外差距更多体现为意识上的差距。美国在数据开放、跨部门共享方面做出了表率，而我国对大数据的价值和应用，政府、学术界、产业界

① 维度、活性等概念将在数据资产章节详细说明，是数据资产评估模型的一部分。
② 许连捷现任中国民间商会副会长，泉州市工商联主席，第十届全国工商联副主席。

和资本市场尚待达成一致的认知。各部门、各地方普遍存在“数据割据”和“数据孤岛”现象，缺乏大数据意识是阻碍我国大数据技术在各行业落地的关键因素。

大数据时代，有两点非常有利于中国信息产业跨越式发展。第一，大数据技术以开源为主，迄今为止，尚未形成绝对技术垄断。即便是 IBM、甲骨文等行业巨擘，也同样是集成了开源技术，与本公司原有产品更好地结合而已。开源技术对任何一个国家都是开放的，中国公司同样可以分享开源的蛋糕。但是需要更加开放的心态、更加开明的思想，正确地对待开源社区。第二，中国人口和经济规模决定中国的数据资产规模冠于全球，客观上为大数据技术的发展提供了演练场。第二点亟待政府、学术界、产业界、资本市场四方通力合作，在确保国家数据安全的前提下，最大程度地开放数据资产，促进数据关联应用，释放大数据的大价值。

目前，政府和产业界积累了大量的数据资产，但是苦于缺乏行之有效的与工程实践匹配的算法和人才来充分挖掘数据的价值。形象地说，好多行业是守着“金山要饭吃”。而学术界，尤其是应用数学领域，在统计学习、图像处理、网络科学领域钻研颇深，但缺乏大量的实际数据来验证和训练算法。虽有屠龙术，无处展身手，两方长期处于脱节的状态。如果应用数学界和产业界紧密协作，将是中国公司的极大利好，会大大促进公司的发展。2012 年 11 月 17 日，在北京大学国际数学研究中心召开了首届“数据科学与信息产业研讨会”。学术界和企业界的一百多位领军人物和活跃分子聚集在一起，共同商讨数据科学的含义和发展计划，以及企业界的需求。这次会议为促进学术界和信息产业的联合，开了一个好头。

数据资产并不是大公司才有的专利。在第七章中将详细讨论一种“泛互联范式”，“终端”+“平台”+“应用”，最后形成数据资产。许许多多富有活力的公司，均符合这一范式。这也是创业型公司开启大数据之路的总结和探索。

自从我们在中国资本市场第一个发出“大数据时代即将到来”的声音后，大数据已经成为年度热词。综合政府、学术界、产业界的最新动向，笔者预计，如果把 2012 年看成大数据普及之年，那么 2013 年将成为大数据应用之年，相关产业规划、行业政策将纷纷出台，金融、电信、政府、电商、医疗、平安城市等相关应用将加

速推进；2014 年至 2016 年将是大数据效益之年，若干中国大数据公司相关业务形成爆发性增长，部分相关公司海内外融资或 IPO 上市。

本书的内容将围绕大数据对产业走势、融合、变迁的影响，在产业中的具体应用（商业模式），以及数据科学的兴起三大主题展开。本章包括大数据产生的历史背景、激动人心的典型特征、系统全面的认知框架等内容，最后会简略说明推广大数据面临的困难和挑战。

第一节　大数据产生的历史背景

提要：

1. 信息基础设施持续完善，包括网络带宽的持续增加、存储设备性价比不断提升，犹如高速公路之于物流，为大数据的存储和传播准备物质基础。
2. 互联网领域的公司最早重视数据资产的价值，最早从大数据中淘金，并且引领大数据的发展趋势。
3. 云计算为大数据的集中管理和分布式访问提供了必要的场所和分享的渠道。大数据是云计算的灵魂和必然的升级方向。
4. 物联网与移动终端持续不断地产生大量数据，并且数据类型丰富，内容鲜活，是大数据重要的来源。

信息科技进步

如果把信息技术的不断进步看成世界万物持续数字化的过程，则会理出一条清晰的主线。信息科技具有三个最核心和基础的能力：信息处理、信息存储和信息传递，几十年来这三个能力的飞速进步，是人类科技史上最为激动人心的故事之一。

1965 年，戈登・摩尔[1]（Gordon Moore）发现芯片上可容纳的晶体管数目，每隔 18 个月左右便会增加一倍，性能也将提升一倍，即摩尔定律。在摩尔定律的指引下，信息产业周期性地推出新的计算机，操作系统和计算能力均在不断提高。工业界和个人都不断地升级计算机设备，从而推动信息产业的巨大进步。每当英特尔开发出计算能力更强的芯片，微软公司就会适时推出功能更强大、操作更方便的操作系统，带动客户新一轮的升级换机热潮。这种循环持续不间断地上演了 40 余年。这段波澜壮阔的历史，使信息处理和存储能力获得成千上万倍的提升。

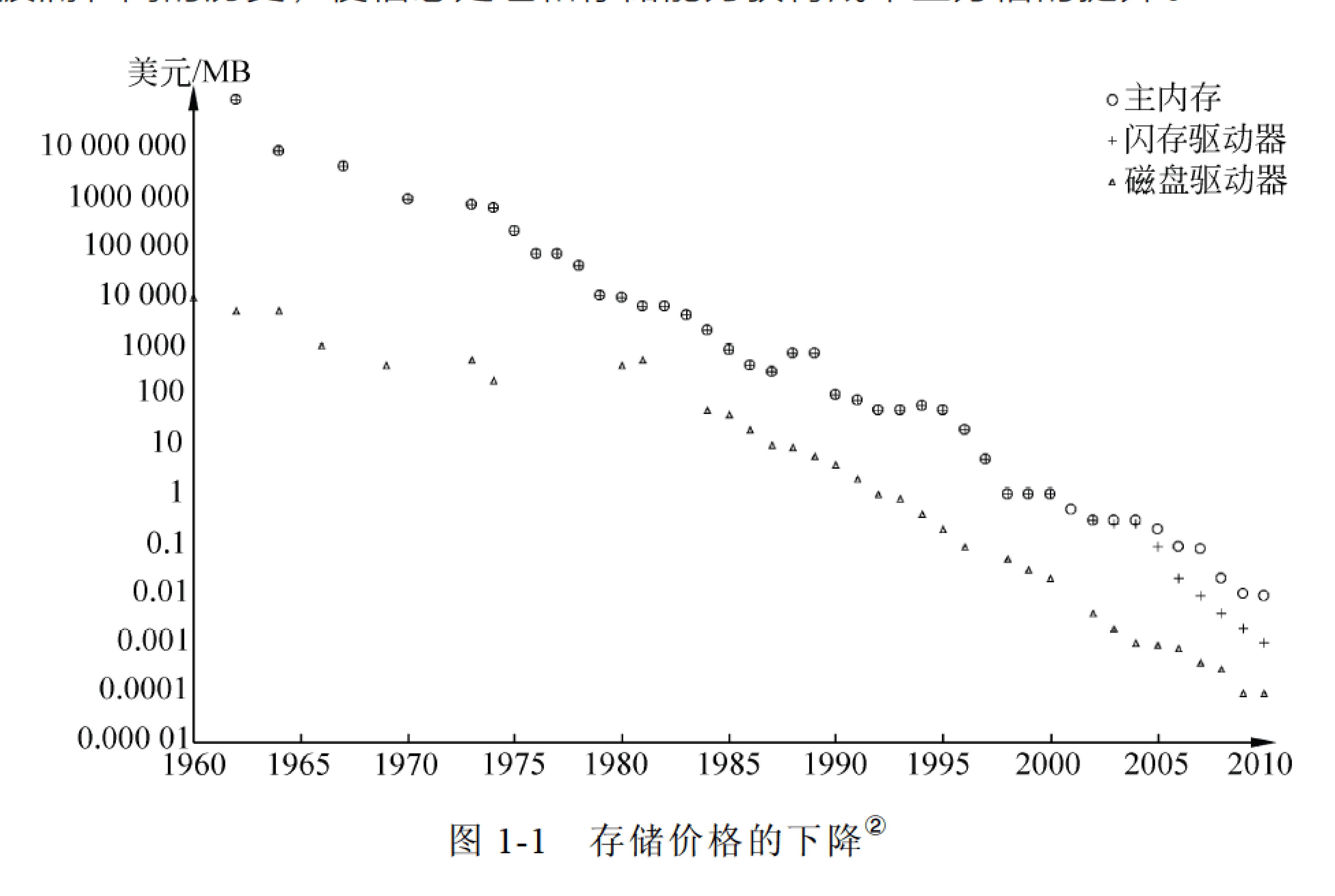


图 1-1　存储价格的下降[2]

1977 年，世界上第一条光纤通信系统在美国芝加哥市投入商用，速率为 45Mbit/s，自此，拉开了信息传输能力大幅跃升的序幕。有人甚至将光纤传输带宽的增长规律称为超摩尔定律，认为带宽的增长速度比芯片性能提升的速度还要快。

---

① 摩尔 1929 年出生在美国加州的旧金山，曾获得加州大学伯克利分校的化学学士学位，并且在加州理工大学（CIT）获得物理化学（physical chemistry)博士学位。20 世纪 50 年代中期，他和集成电路的发明者罗伯特・诺伊斯（Robert Noyce）一起在威廉・肖克利半导体公司工作。1968 年，摩尔和诺伊斯创办了大名鼎鼎的英特尔公司。自 1982 年起的 10 年间，微电子技术共有 22 项重大突破，其中由（英特尔）公司开发的就有 16 项之多。摩尔在 1974 年至 1987 年间担任英特尔公司的总裁和首席执行官，英特尔公司在微机时代和微软公司一道主宰了整个产业的发展。

② 来源：Plattner and Zeier, “In-Memory Data Management”, 2011, p. 15-16; * Driscoll, “Big Data Now”。

事实上，存储的价格从 20 世纪 60 年代 1 万美元 1MB，降到现在的 1 美分 1GB 的水平，其价差高达亿倍，如图 1-1 所示。在线实时观看高清电影，在几年前还是难以想象的，现在却变得已习以为常了。网络的接入方式也从有线连接向高速无线连接的方式转变。毫无疑问，网络带宽和大规模存储技术的高速持续发展，为大数据时代提供了廉价的存储和传输服务，如图 1-2 所示。因而，本书假定存储和带宽不再是制约数据应用的因素。

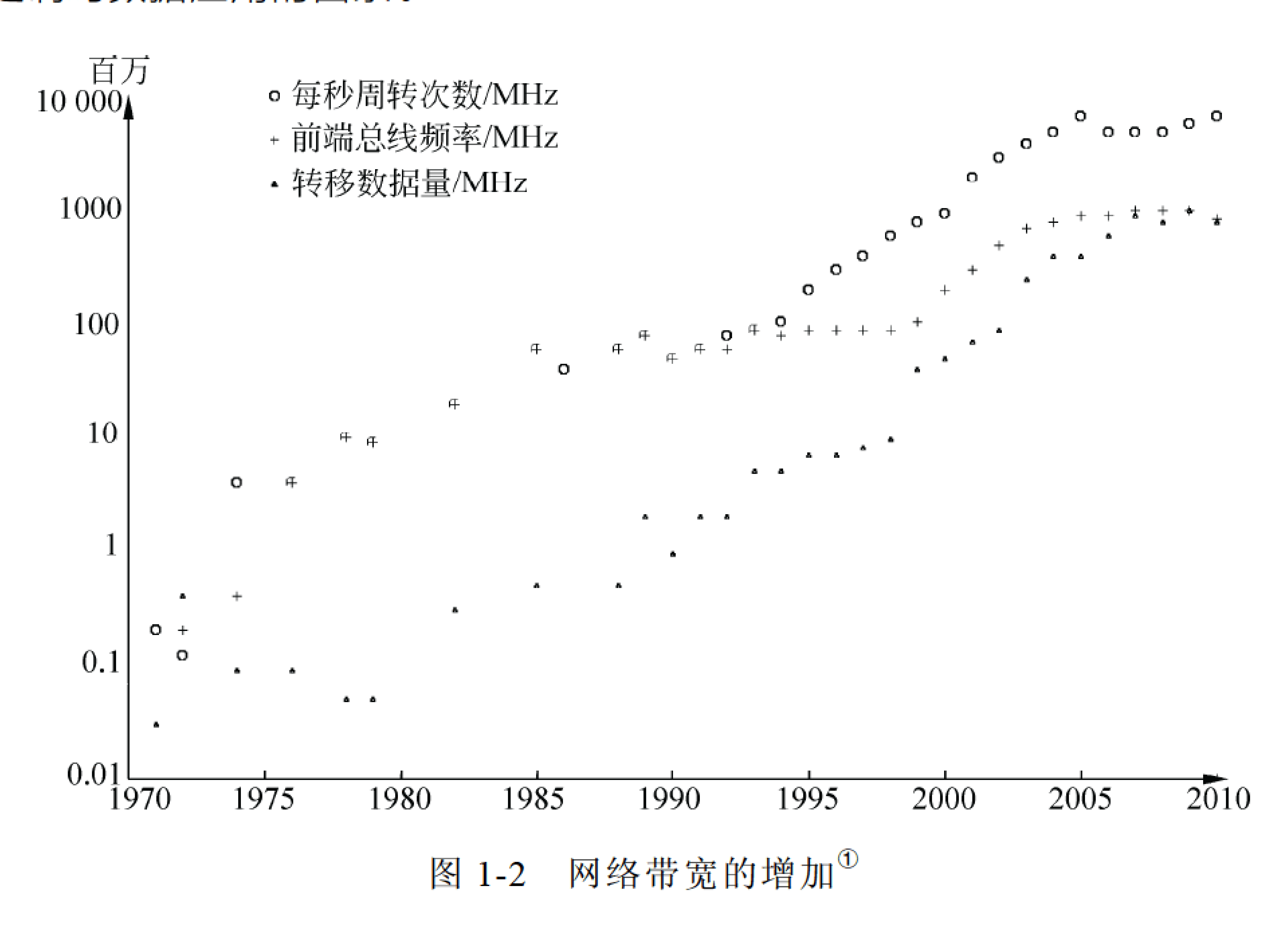


图 1-2　网络带宽的增加[①]

互联网的诞生

互联网的出现，在科技史上可以比肩“火”与“电”的发明。这个伟大的发明同样是由军事目的驱动的。计算机在军方应用得越广泛，计算机上保存的军事机密就越多。人们担心如果保存重要军事机密数据的主要计算机被摧毁的话，很可能就会输掉整个战争，于是，推动计算机之间互相传递数据并互为备份的通信机制被提上日程。1969 年，把分属于不同大学的四台计算机互相连接起来，这就是最早的互

① 来源：Plattner and Zeier, “In-Memory Data Management”, 2011, p. 15-16; * Driscoll, “Big Data Now”

联网雏形。

互联网把每个人桌面上的计算机连接起来，改变了人们的生活，成为大家获取各类数据的首要渠道。通过互联网获取数据的模式可以简单地抽象为“请求”+“响应”的模式。理解这种获取信息的方式，有助于理解“大数据”的价值，所以笔者多花些笔墨把这个模式解释清楚。

互联网上的“脚印”

用收音机听广播，或者用电视机看电视节目，都是“广播”+“接收”的模式。不管有没有电视机在接收信号，广播塔总是在发送电视节目信号。随时打开电视机，随时就能收看电视节目。在“广播”+“接收”模式中，广播塔是不知道有谁在接收节目的，如图 1-3 所示。

“请求”+“响应”模式则不同，如果客户端（所有接入互联网的设备、软件等）不主动要求，服务器端是不会发送任何数据的，如图 1-4 所示。互联网应用协议基本上都是这种模式。当然也有“广播”+“接收”模式的协议，但是不常用。每一次访问请求其实就是一次鼠标点击操作，服务器的日志中，忠实地记录下来每个人访问的时间、请求的命令、访问的网址等数据。这些访问记录就像人们在雪地上行走留下的脚印一样，“脚印”连成一串，构成了人们在互联网上的“行为轨迹”。想一想猎人是怎样通过追踪脚印捕获猎物的，就会明白这些“轨迹”中蕴含着巨大的价值。所以，各类服务器上的日志就是一种非常重要的大数据类型。

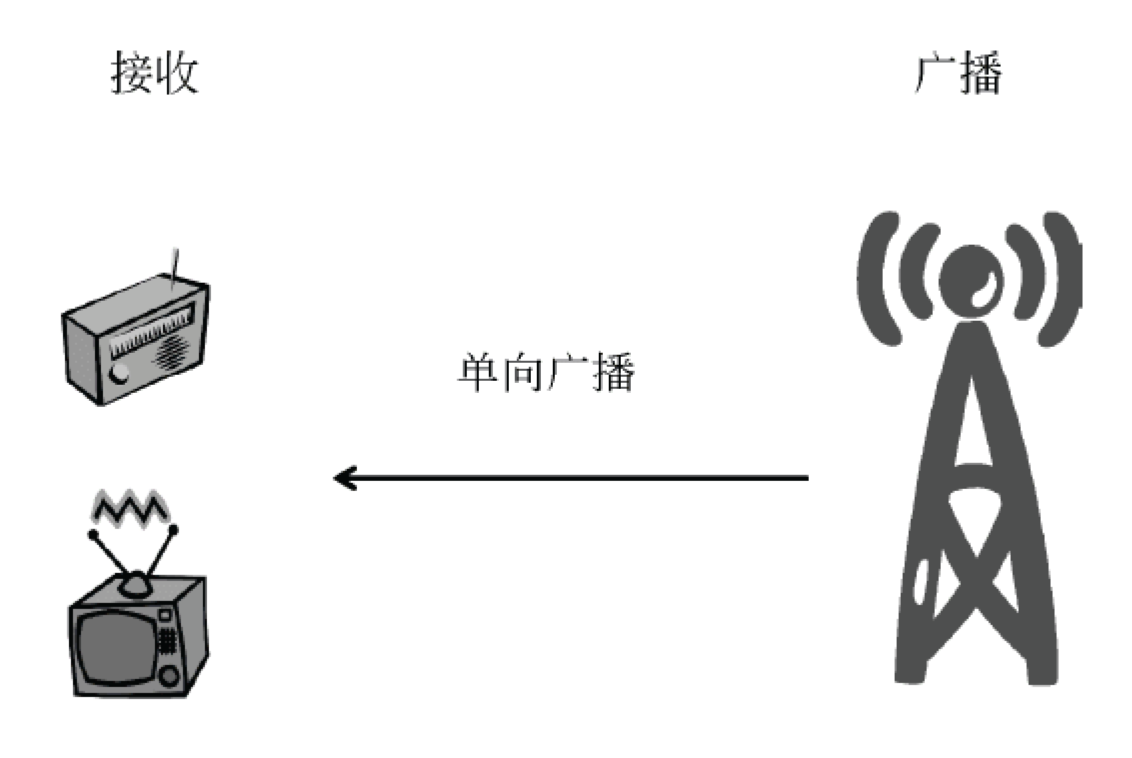


图 1-3 “广播”+“接收”模式

曾经有制作服装的公司想要调查顾客的购买意愿。需要统计顾客拿起了哪件衣服？试穿了哪件衣服？在专卖店逗留了多长时间？这就需要安装摄像头，选样本，可能花费上亿的资金。要想省钱的话，其结果可能会失去参考价值。如果在网上做同样的事情，成本近乎为“0”。大家可以想想，在淘宝网或者京东商城的主页上，每一个网页都相当于一家店铺，打开这个网页就等于进入了店铺；点击了衣服，相当于顾客拿起衣服仔细端详；把衣服放到收藏夹，可以理解为试穿。这样，在实体店中顾客的行为几乎被完整地映射到网页上了。不同的是，互联网忠实地记录下“顾客”在“店”里停留的时间、关心的品类；此外，顾客和销售员的对话、顾客与顾客之间的对话，也被忠实地记录、保存。互联网企业做与那家制衣公司同样的调查，成本近乎为“0”。

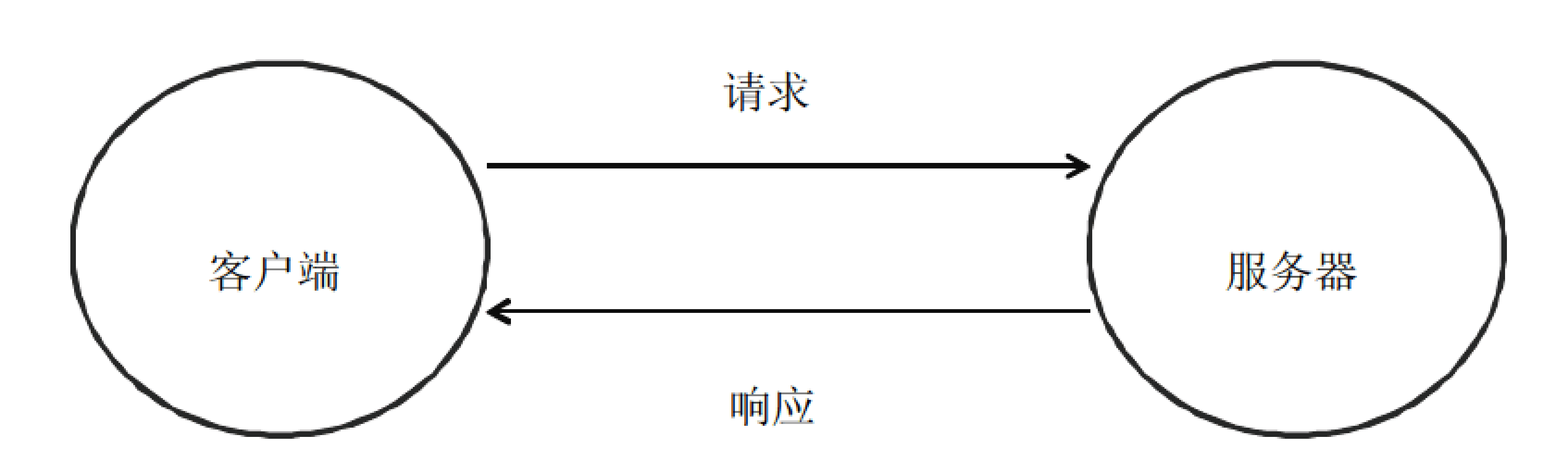


图 1-4 “请求”+“响应”模式，记录用户的请求

因为互联网的内在机理，使互联网成为大规模接近消费者、最理解消费者的工具和平台。互联网没有删除键，人们在互联网上的一言一行都被忠实地记录。古代皇帝身边总有一位兢兢业业的史官，随身携带纸笔，记下皇帝的起居作息、金口玉言。互联网就像每个人的“史官”，它从不知疲倦，事不分大小，悉心而精准地记录着一切。事实上，这位“史官”记录的就是大家的数字化生活，如图 1-5 所示。

云计算与大数据

云计算，再一次改变了数据的存储和访问方式。在云计算出现之前，数据大多分散保存在每个人的个人计算机中、每家企业的服务器中。云计算，尤其是公用云

计算，把所有的数据集中存储到“数据中心”，也即所谓的“云端”，用户通过浏览器或者专用应用程序来访问。

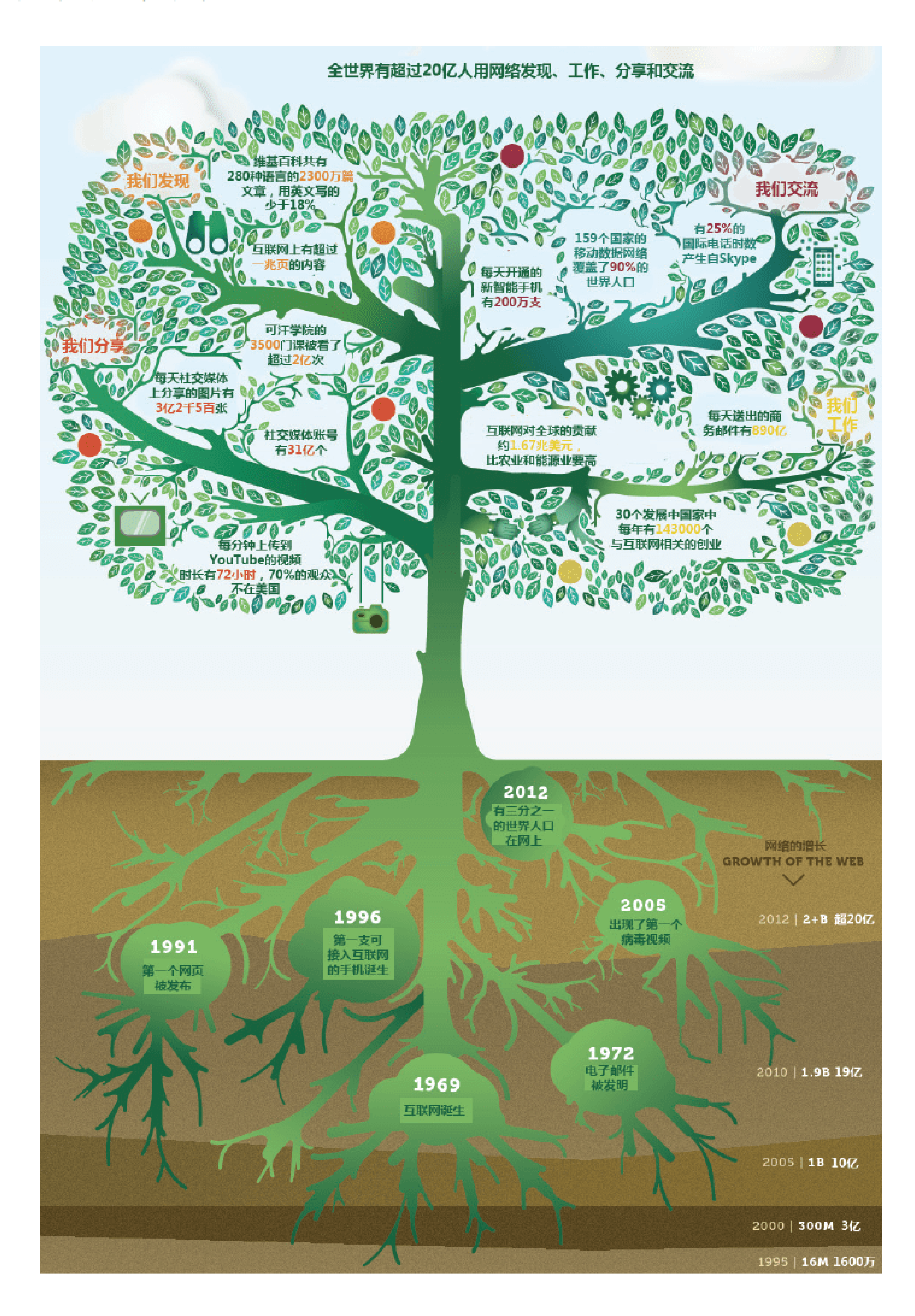


图 1-5 网络生活（来源：谷歌）

一些大型的网站，通过提供基于“云”的服务，积累大量的数据，成为事实上的“数据中心”。“数据”是这些大型网站最为核心的资产。他们不惜花费高昂的费用付出巨大的努力来保管这些数据，以便加快用户的访问速度。谷歌公司甚至购买了单独的水力发电站，为其庞大的数据中心提供充足的电力。根据一些公开资料显示，谷歌在全球分布着 36 个数据中心。图 1-6 是谷歌公司数据中心内一景，大家

可以由此领略到科技之美。

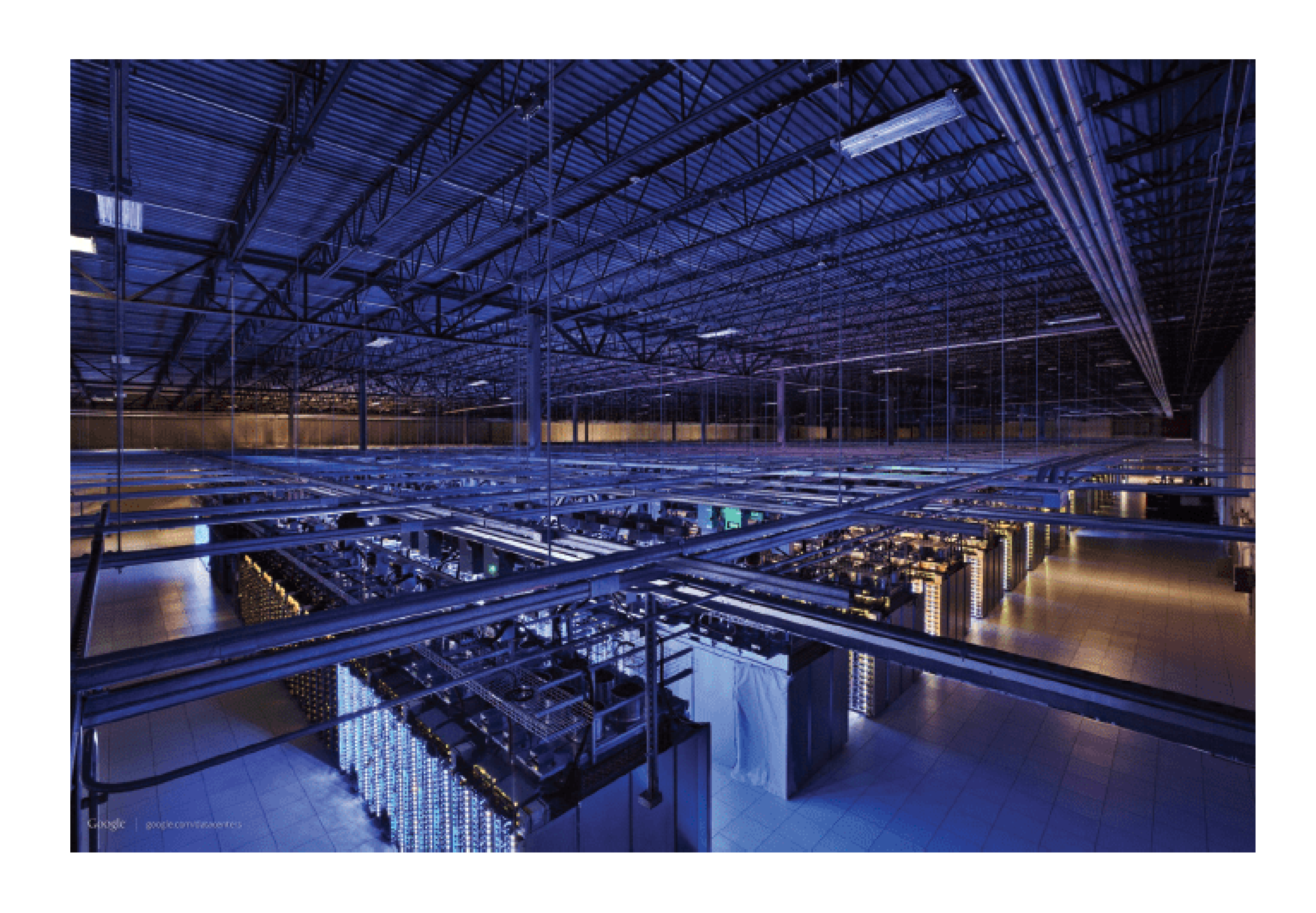


图 1-6　谷歌数据中心一景（来源：谷歌）

这几年国内各地兴起了建设云计算基地的风潮，客观上为“大数据”的诞生准备了必备的储存空间和访问渠道。各大银行、电信运营商、大型互联网公司、政府各个部委都拥有各自的“数据中心”。银行、电信、互联网公司绝大部分已经实现了全国级的数据集中工作。

在笔者的大数据报告中曾经提了一个观点，引起了广泛的关注和争议：“没有大数据的云计算，就是房地产的代名词[①]”。云计算确实可以称为一场信息技术领域内的革命，甚至对社会也必将产生革命性的影响，但是它却并不是一场技术革命，云计算在本质上是一场 IT 产品/服务消费方式的变革，云计算中的一个广为宣传的核心技术——虚拟化软件，早在 20 世纪 60 年代就已经被应用在 IBM 的大型主机中了。

云计算是大数据诞生的前提和必要条件。没有云计算，就缺少了集中采集和存储数据的商业基础。云计算为大数据提供了存储空间和访问渠道；大数据则是云计算的灵魂和必然的升级方向。

2012 年，业内所有的云计算大会，无论官方背景还是民间主办，都是把“大数

① 参见国金证券大数据系列研究报告第一篇《大数据时代即将到来》，第 14 页。

据”作为一个核心的主题。甚至有时候都分不清楚，这是云计算的会，还是大数据的会。

物联网

物联网是另一个信息技术领域的热词，究其本质是传感器技术进步的产物。遍布大街小巷的摄像头，是大家可以直观感受到的一种物联网形态。事实上，传感器几乎无处不在，使用它可以监测大气的温度、压强、风力，监测桥梁、矿井的安全，监测飞机、汽车的行驶状态。一架军用战斗机上的传感器多达数千个。现在大家常用的智能手机中，就包括重力感应器、加速度感应器、距离感应器、光线感应器、陀螺仪、电子罗盘、摄像头等各类传感器。这些不同类型的传感器，无时无刻不在产生大量的数据。其中的某些数据被持续地收集起来，成为大数据的重要来源之一。

社交网络

社交网络是互联网发展史上的又一个重要的里程碑。它把人类真实的人际关系完美地映射到互联网空间，并借助互联网的特性而大大升华。广义的看，社交网络使得互联网甚至具备某些人类的特质，譬如“情绪”：人们分享各自的喜怒哀乐，并相互传染传播。社交网络为大数据带来一类最具活力的数据类型，人们的喜好和偏爱。更重要的是，人们还知道在社交网络中，如何利用网民的关系链来传播这些喜好和偏爱。这就为研究消费者行为打开了另一扇方便之门。如果深入地分析社交网络就会发现，大型的社交网络平台事实上构成了以“个人”为枢纽的不同的数据的集合。借助“分享”按钮，人们在不同网站上的购物信息、浏览的网页都可以“分享”在社交网络上。想想前面提到的雪地上的脚印，社交网络把网民在不同网站上留下的“脚印”链接起来，形成完整的行为轨迹和“偏好”链。

图 1-7 是 Facebook 的一个实习生把网站中人们相互联系的数据通过建模、渲染得到的一幅图片，越是明亮的地方，人们相互交流越是活跃。现在 Facebook 是

世界上最大的社交网站，每月的活跃用户数突破了 10 亿。

智能终端普及

古人只能用“大漠孤烟直，长河落日圆”等诗词歌赋来主观描述他们的所见所闻，我们则可以掏出手机、照相机、摄像机，再现美丽的风景，与亲朋好友分享。执着的古人迷路时索性信马由缰不问归路[①]，我们则可以拿出智能手机使用导航软件找到目的地。

智能终端不仅仅局限于个人应用，许多行业都已经开始大规模地部署终端产品。举一个“美丽”的例子，婚纱摄影行业：以前影楼需要租用大面积的场馆、位置优良租金高昂的门店，携带大型笨重的写真集，展示给准新娘们用以挑选照片。但是如今利用 iPad，可以做出令人心醉神迷的实景效果，如 360° 旋转等特效。准新娘只需要一部 iPad，就可以全面地看到最终的拍摄效果，并利用其交互特性提高样片选择的精准度。



图 1-7　反映社交网络 Facebook 上人们活跃程度的世界地图（来源：Facebook）

KPCB[②]（凯鹏华盈）是美国最大的风险投资基金之一，其合伙人 Mary Meeker

---

① 《晋书・阮籍传》中记载，“时率意独驾，不由径路，车迹所穷，辄恸哭而反”。籍非迷路，刻意为之。正文中是夸张的说法。

② KPCB（Kleiner Perkins Caufield & Byers）公司成立于 1972 年，是美国最大的风险投资基

在 2012 年发布的一份趋势报告中指出，在 2010 年第二季度，智能手机加平板电脑的出货量已经超越台式机和传统笔记本电脑（参见图 1-8），并且预计在 2013 年第二季度，智能移动终端全球保有量也将实现反超（参见图 1-9①）。

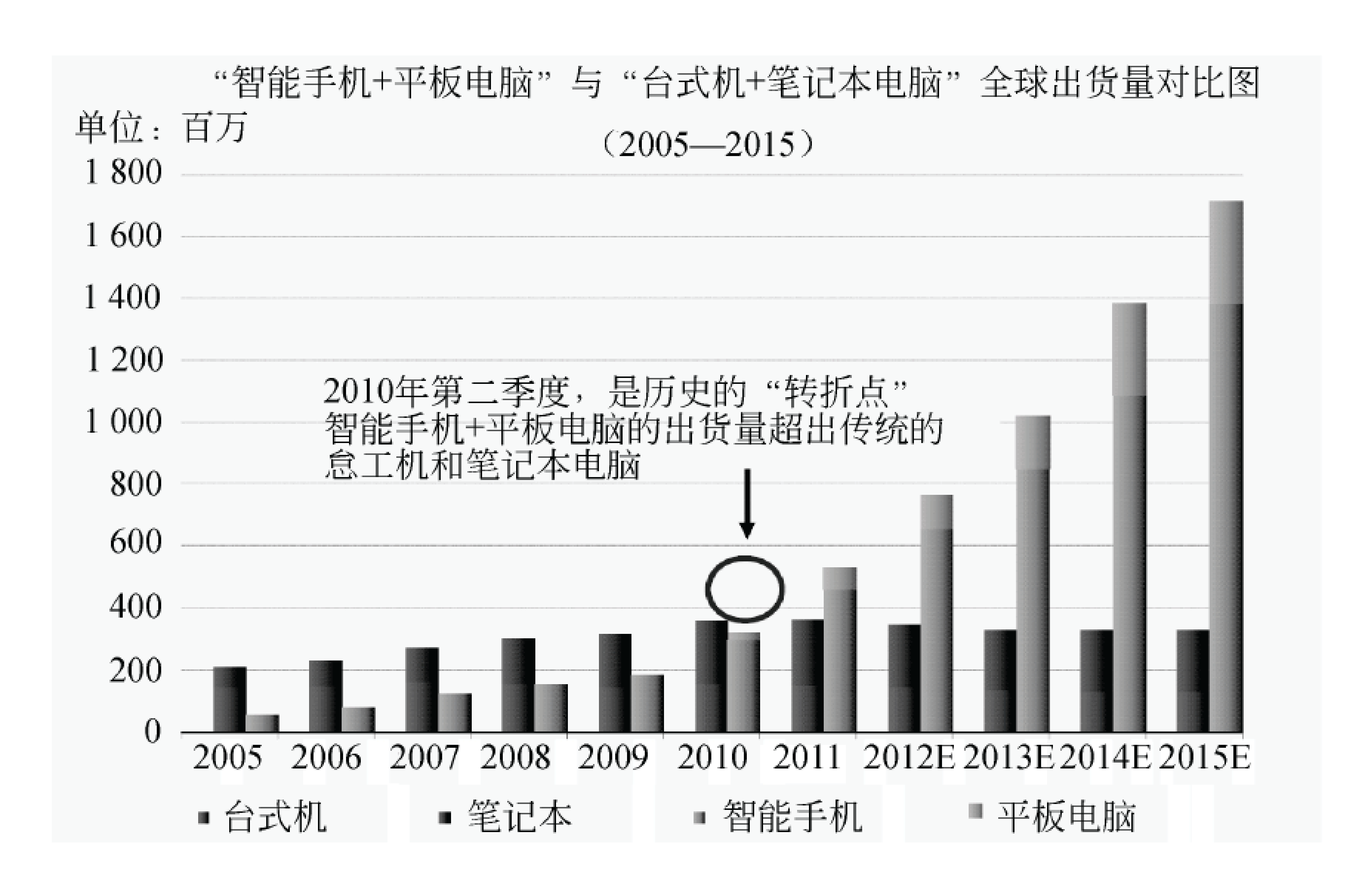


图 1-8　移动设备与传统台式机、笔记本电脑的全球出货量对比图（来源：Katy Huberty, Ehud Gelblum, Morgan Stanley Research. Data and Estimates as of 9/12.）

智能终端的普及给大数据带来了丰富、鲜活的数据。苹果公司 2012 年公布的一组运营数据可以反映智能终端上人们的活跃程度。其中，iMessage 功能目前每秒为用户传递 28 000 条信息；iCloud 已经为用户提供了总计 1 亿多份的文档；GameCenter 的账号创建数达到了 1.6 亿，当前 iOS 应用总数突破了 70 万，支持 iPad 的应用则达到了 27.5 万；苹果 AppStore 的应用下载量突破了 350 亿次大关，通过分成付给应用开发商的分成总额已达 65 亿美元；iBooks 中的图书总数已达 150 万册，下载量也超过了 4 亿。

金之一，主要是承担各大名校的校产投资业务。KPCB 公司人才济济，在风险投资业首屈一指，在其所投资的风险企业中，有康柏公司、太阳微系统公司、莲花公司等计算机及软件行业的佼佼者。随着互联网的飞速发展，公司抓住这百年难觅的商业机遇，将风险投资的重点放在互联网产业上，先后投资美国在线、奋扬（EXICITE）、亚马逊书店、网景、谷歌、Intuit 等公司。

① 计算保有量，预计保有量，假定台式机的换机周期是 5 年，笔记本电脑的换机周期是 4 年，智能手机是 2 年，平板电脑是 2.5 年。



图 1-9 移动设备与传统台式机、笔记本电脑的全球保有量对比图（来源：Katy Huberty, Ehud Gelblum, Morgan Stanley Research. Data and Estimates as of 9/12.）

第二节 大数据的定义和特征

提要：

1. 未来的不确定性是人类产生恐惧的根源之一，也是各类组织最为头痛的问题。大数据技术让人们看到了解决未来预测问题的一丝曙光。
2. 大数据四个典型的特征：第一，数据量巨大；第二，数据类型多样；第三，数据中富含价值；第四，必须在尽可能短的时间内发掘出价值。
3. 尽管本节重点介绍大数据的四个特征，但是并非只有数据量大才能称为大数据。人们更看重的是"快速地从各类数据中获得信息的能力"。

麦肯锡（美国首屈一指的咨询公司）是研究大数据的先驱。在其报告《Big data: The next frontier for innovation, competition,and productivity》中给出的大数据

定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据[①]。

国际数据公司（IDC）从大数据的四个特征来定义，即海量的数据规模（Volume）、快速的数据流转和动态的数据体系（Velocity）、多样的数据类型（Variety）、巨大的数据价值（Value）。

亚马逊（全球最大的电子商务公司）的大数据科学家 John Rauser 给出了一个简单的定义：大数据是任何超过了一台计算机处理能力的数据量。

维基百科中只有短短的一句话："巨量资料(big data)，或称大数据，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。"

大数据是一个宽泛的概念，见仁见智。上面几个定义，无一例外地都突出了"大"字。诚然"大"是大数据的一个重要特征，但远远不是全部。笔者在调研多个行业后，给出了自己的定义：大数据是"在多样的或者大量数据中，迅速获取信息的能力[②]"。前面几个定义都是从大数据本身出发，我们的定义更关心大数据的功用，它能帮助大家干什么。在这个定义中，重心是"能力"。大数据的核心能力，是发现规律和预测未来。

发现规律，预测未来

任何行为，皆有前兆。但在现实世界中，缺少实时记录的工具，许多行为看起来是"人似秋鸿有来信，事如春梦了无痕"。在互联网世界则完全不同，是"处处行迹处处痕"。要买商品，必先浏览、对比、询价；要搞活动，必先征集、讨论、策划。互联网的"请求"+"响应"机制恰恰在服务器上保留了人们大量的前兆性的行为数据，把这些数据搜集起来，进一步分析挖掘，就可以发现隐藏在大量细节背后的规

① 参见麦肯锡，《Big data: The next frontier for innovation, competition, and productivity》，2011 年。

② 参见国金证券大数据系列研究报告第二篇《大数据时代的三大发展趋势和投资方向》，第 7 页。

律，依据规律，预测未来。收集分析海量的各种类型的数据，并快速获取影响未来的信息的能力，就是大数据技术的魅力所在。

1993 年，《纽约客》刊登了一幅漫画，标题是“互联网上，没有人知道你是一条狗”，如图 1-10 所示。据说作者彼得·施泰纳因为此漫画的重印而赚取了超过 5 万美元。当时关注互联网社会学的一些专家，甚至担忧“计算机异性扮装”而引发的社会问题。譬如，同性恋和恋童癖可能会借助互联网而大行其道。

20 年后，互联网发生了巨大的变化，移动互联、社交网络、电子商务大大拓展了互联网的疆界和应用领域。人们在享受便利的同时，也无偿贡献了自己的“行踪”。现在互联网不但知道对面是一条狗，还知道这条狗喜欢什么食物、几点出去遛弯、几点回窝睡觉。人们不得不接受这个现实，每个人在互联网进入到大数据时代都将是透明性存在的。

事实上，对于未来的不确定性是人类产生恐惧的根源之一，也是各类组织最为头痛的问题。大数据技术让人们看到了解决未来预测问题的一丝曙光。通过利用大数据技术，可以预测自然天气的变化，预测个体未来的行为，甚至预测某些社会事件的发生。它会让人们的生活更为从容，让决策不再盲目，让社会更加高效的运转。这就是大数据技术带给人们的好处。全球复杂网络权威巴拉巴西认为，人类行为 93%是可以预测的。笔者的确不知道这位老先生是怎么计算出来 93%这个数字的，但大数据可以预测未来是显而易见的，这是首个使人类具备了预测短期未来的技术。

听起来似乎很玄妙，大数据不就是算命先生么？

其实，或多或少，人们都具备预测的能力。譬如，儿子跟小伙伴们疯玩，我知道他肯定在 7 点之前会回家，因为他饿了。再如，家乡流传的很多谚语，其中一句“八月十五云遮月，正月十五雪打灯”，说明大自然就有许多规律性的东西。估计现在的科学也没有办法解释几乎半年跨度内气象间的因果关系，但是几千年的观察和积累却发现了它。自然、社会、商业无不服从某些规律，大国兴衰、王朝更替亦有规律可循。只是过去囿于技术条件人们无法记录下造成某件事情发生的先兆数据，

无法去计算其中的因果关系。这些规律要么被神秘化，要么被庸俗化。



“On the Internet, nobody knows you’re a dog.”

图 1-10　“互联网上，没有人知道你是一条狗”（来源：www.chrisabraham.com）

任何事情的发生，都会有蛛丝马迹的前兆表露出来。如果人们不去关注一支股票的行情走势，就不会去买卖这支股票；如果人们从不去询问某件商品的价格，也很难产生购买行为；如果事先没有联络沟通，人们就很难聚在一起；如果没有闷热的天气，似乎就没有透心凉的大雨。关于地震前种种异象，更是被许多书籍、文章大肆渲染。

假定有一种技术可以记录下所有这些先兆，人们就获得了未卜先知的能力。利用大数据技术，能够广泛采集各种各样的数据类型，并进行统计分析，从而预测未来。大数据影响之深远，波及之广泛，远非一般的信息技术可比。

“过去我认为我的工作就是追捕罪犯，而现在对这项工作有了全新的认识，我们

分析犯罪数据，识别犯罪模式，并部署警力，帮助美国部分城市重大犯罪率降低了30%。终结犯罪，在案发之前。”这是 IBM 公司的一则广告，宣传利用大数据构建智慧的地球。

“2008 年初，阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购在下滑。海关是卖了货出去以后再获得数据，而我们提前半年时间从询盘上就推断出世界贸易发生变化了。”通常而言，买家在采购商品前，会比较多家供应商的产品，反映到阿里巴巴网站统计数据中，就是查询点击的数量和购买点击的数量会维持一个相对的比例。统计历史上所有买家、卖家的询价和成交数据，可以形成询盘指数和成交指数，这两个指数是强相关的。询盘指数是前兆性的，前期询盘指数活跃，就会保证后期一定的成交量。所以，当马云观察到询盘指数异乎寻常的下降后，自然就可以推测未来成交量的萎缩。这种统计和分析，如果缺少大数据技术的支持，是难以完成的。这次事件，马云提前呼吁、帮助成千上万的中小制造商准备过冬粮，从而赢得了崇高的声誉。

中国建设银行的电子商务金融平台——“善融商务”于 2012 年 6 月 28 日正式开业。官方的宣传：“善融商务是建设银行顺应电子商务发展潮流，结合传统金融服务优势和新兴电子商务服务应用而搭建的全流程、综合性的电子商务服务平台。”据说建行内部推进电子商务的力度非常大，分行考核严厉，甚至亏本也要把小商家搬到网上。银行建立电子商务交易平台，听起来像不务正业，其实是醉翁之意不在酒。银行需要那些小商家的经营数据，来预测商家的贷款需求和还款能力，从而大幅降低小额借贷风险。建行此举，不论成功与否，都足以证明建行高层深刻地理解了大数据的重要性和其惊人的预测能力。这种能力，对建行而言，就意味着低风险、高收益，是每家金融机构都梦寐以求的境界。常常说富贵险中求，传统经营一般是高风险、高收益，不料有了大数据在手，就能低风险、高收益，难怪金融机构趋之若鹜。如果金融机构再不重视大数据的潜在价值，行将成为 21 世纪的恐龙，不复往日的荣光。

数据大爆炸

截至 2011 年，全球拥有互联网用户数已达到 20 亿； RFID 标签在 2005 年的保有量仅有 13 亿个，但是到 2010 年这个数字超过了 300 亿；2006 年资本市场的数据比 2003 年增长了 17.5 倍；目前新浪微博上每天上传的微博数超过 1 亿条；Facebook 每天处理 10TB 的数据；世界气象中心积累了 220TB 的 Web 数据，9PB 其他类型数据……

根据国际数据公司（IDC）的《数据宇宙》报告显示：2008 年全球数据量为 0.5ZB，2010 年为 1.2ZB，人类正式进入 ZB 时代。更为惊人的是，2020 年以前全球数据量仍将保持每年 40%多的高速增长，大约每两年就翻一倍，这与 IT 界人尽皆知的摩尔定律极为相似，姑且可以称之为“大数据爆炸定律”。预计 2015 年全球数据量将达到 7.9ZB，2020 年将突破 35ZB，是 2008 年的 70 倍、2011 年的 29 倍，如图 1-11 所示。

同时，根据互联网数据中心的《中国互联网市场洞见：互联网大数据技术创新研究 2012》报告显示：截至 2011 年年底，中国互联网行业持有的数据总量已达到 1.9EB，预计 2015 年该规模将增长到 8.2EB 以上。

人类社会的数据量在不断刷新一个个新的量级单位，已经从 TB、PB 级别跃升至 EB、ZB 级别。然而，35ZB、8.2EB 究竟是一个什么样的概念呢？为此，首先了解下面几组关于数据衡量单位的公式：

1B = 8 bit

1KB = 1024B ≈1000 byte

1MB = 1024 KB ≈1 000 000 byte

1GB = 1024 MB ≈1 000 000 000 byte

1TB = 1024 GB ≈1 000 000 000 000 byte

1PB = 1024 TB ≈1 000 000 000 000 000 byte

1EB = 1024 PB ≈1 000 000 000 000 000 000 byte

1ZB = 1024 EB ≈1 000 000 000 000 000 000 000 byte

1YB = 1024 ZB ≈1 000 000 000 000 000 000 000 000 byte

一本《红楼梦》共有 87 万字(含标点),每个汉字占两个字节,即 1 个汉字=2B,由此计算 1EB 约等于 6626 亿部红楼梦。美国国会图书馆是美国四个官方图书馆之一，也是全球最重要的图书馆之一，截至 2011 年 4 月，藏书约为 1.5 亿册，收录数据 235TB，1EB 约等于 4462 个美国国会图书馆的数据存储量。

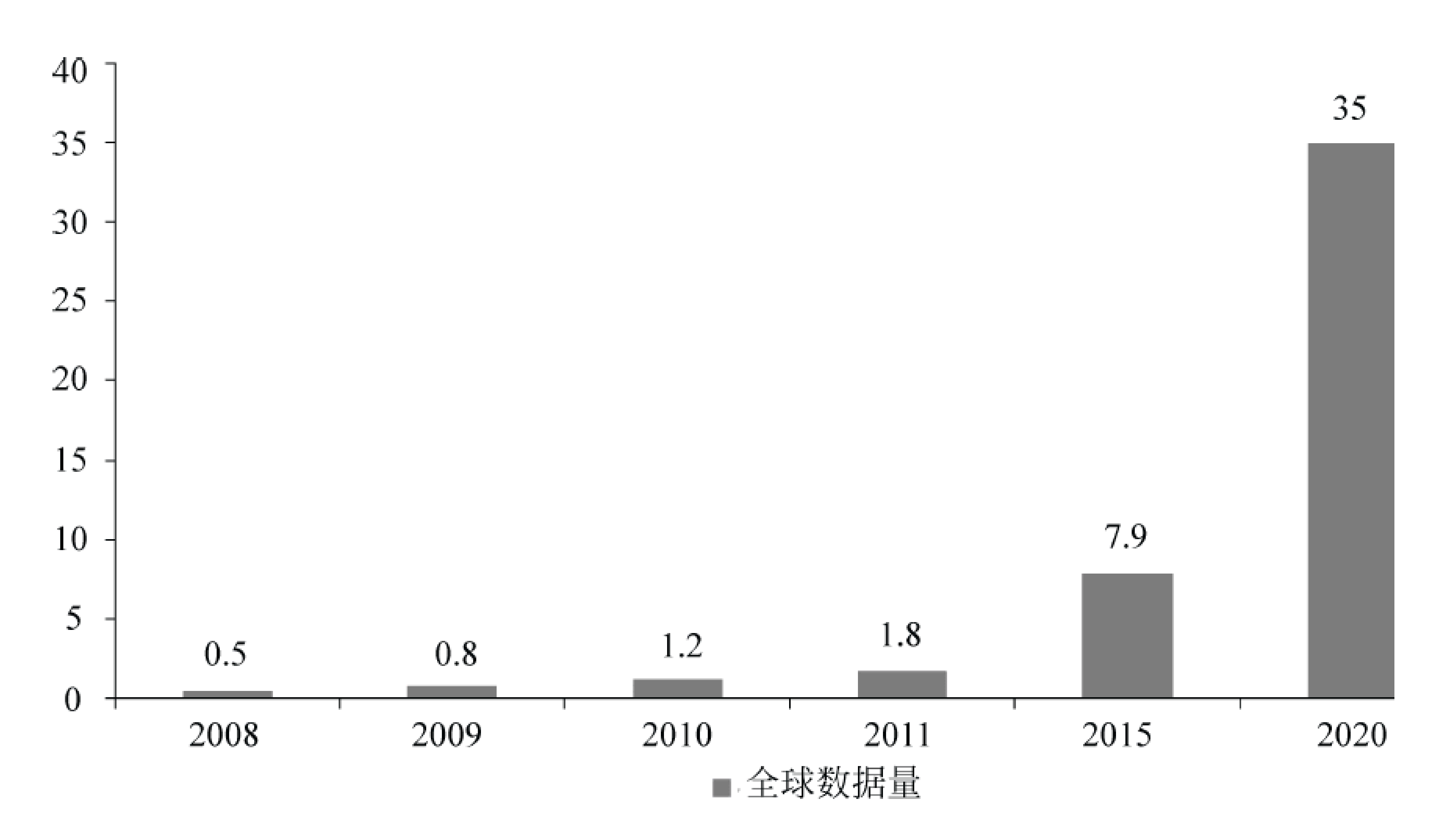


图 1-11　全球数据量增长预测（单位：ZB）（来源：IDC 数字宇宙研究）

数据的多样化

电影《黑客帝国》中，主人公尼奥吞下那颗蓝色的小药丸后，发现原来他生活中一切的一切，包括他的工作、伙伴，高楼大厦、天空、大地，甚至喜、怒、哀、乐，都是数字化的幻像。真实的物理世界当然不像电影那样天马行空，但在许多领域的确朝高度数字化的方向演进。

譬如，那些高楼大厦，利用三维建模技术，形成了包含设计、施工、维护等综合的建筑信息模型。在消费者眼中，建筑信息模型呈现出来漂亮、壮观，让人们不

得不掏出钱来买单的效果图；在房地产商老板眼中，建筑信息模型则清楚地告诉他们整个过程应该花多少钱；在设计师眼中，这个模型就是各种各样的设计图的综合，他们可以方便地调整管线走向、通风的设计等；而在工人眼中，建筑信息模型就是施工图；消防部门不用等到完工，通过建筑信息模型就能评估建筑的消防效果和做了人群疏散的动态模拟。也就是说，建设一栋大楼的方方面面都可以是数字化的。

人们日常工作中接触的文件、照片、视频，都包含大量的数据，蕴含大量的信息。这一类数据有一个共同的特点，大小、内容、格式、用途可能都完全不一样。以最常见的 Word 文档为例，最简单的 Word 文档可能只有寥寥几行文字，但也可以混合编辑图片、音乐等内容，成为一份多媒体的文件，来增强文章的感染力。这类数据通常称为非结构化数据。

与之相对应的另一类数据，就是结构化数据。这类数据大家可以简单地理解成表格里的数据，每一条的结构都相同。大家每月都能领到工资条，每个工资条结构都是一样的，当然里面的工资和缴纳的个税、保险不同。每个人的工资条依次排列到一起，就形成了工资表。利用计算机处理结构化数据的技术比较成熟，从事会计、审计等工作的人，利用 Excel 工具很容易进行加减乘除、汇总、统计之类的运算。如果进行大量的运算，一些商业数据库软件就会派上用场，它们专门用于存储和处理这些结构化的数据。

但不幸的是，企业中和人们日常接触到的数据绝大部分都是非结构化的。有的咨询机构认为非结构化数据占企业总数据量的 80%，也有机构认为占 95%，总之，没有权威、准确的统计。如何像处理结构化数据那样，方便、快捷地处理非结构化数据，是信息产业一直以来的努力方向之一。在这个领域，信息业是走了不少弯路的。起初人们借助结构化数据处理的成果，把非结构化数据也用传统的数据库（基于关系型的数据库）来处理。非结构化数据的一大特点就是“龙生九子，各各不同”，硬要套到一个模子里面来，结果是费力不讨好。于是，人们一度认为大量的非结构

化数据是难以处理的。

幸运的是，谷歌公司在为公众提供页面搜索服务的同时，顺便解决了大量网页、文档这类数据的快速访问难题，成为大数据技术的先驱。雅虎公司的一个开发小组，利用谷歌的成果成功地开发出大数据处理的一套程序框架，这就是众所周知的Hadoop。目前，这个领域非常活跃，发展可谓日新月异。

这些公司的实践，让大家面对其他各类的非结构化数据处理难题重建信心，如高清图像、视频、音频等的处理技术都已驶入了快车道。

另外，社交网络上的表现人们情绪的数据日益丰富。例如，[笑脸]、[鼓掌]、[握手]、[愤怒]、[纪念]等代表人们心情的标准化图释的大量使用，无疑表达了人们对某一事件的总体情绪，可能昭示线下会发生某些行为。

大数据的价值特征

7 • 21 北京暴雨之夜，微博成了救灾的明星。一些好心人在微博上公开自己公司地址，方便大家去躲雨和休息。大家依据微博实时了解哪个地方出现了拥堵，哪个地方需要救援。当然，救灾不力应对失当是另外一回事儿。短信、电话都难以描述精确的地址，尤其是当人们焦虑和着急的时候，但是一条微博中可以同时包括人物、时间、地点三个要素，打开微博附加的坐标数据，就可以在地图上迅速定位，为及时救灾提供了方便。在这个例子中，人们看到融合数据的价值。

再如视频监控的例子，银行、地铁等一些敏感的部门或者地点，摄像头都是 24 小时运转，会产生大量视频数据。一般情况下，这些视频数据非常枯燥、乏味，并不会引人注目。但是如果恰巧拍到有图谋不轨的人，那么这一帧图像对公安人员来讲，就是非常有价值的了。然而，人们无法在事前知道哪一帧会有用，只好把所有的视频数据都保存下来，甚至保存了一年的数据，只有那一秒对破案有用。但是在研究人类行为的社会学家眼中，这些视频可能就是难得的第一手资料，也许可以借

此窥探人类的某些行为模式。

笔者曾经读过一篇日本的短篇小说，其情节惊悚。一位年轻貌美却家境贫寒的姑娘，有幸得到一份高薪的工作，照顾一个垂死的病人。奇怪的是，院长要求姑娘必须每时每刻都穿着一件电子背心。医院大楼空空荡荡，令人害怕。姑娘为了养家，不得不忍受大楼里每晚都发生的恐怖事件。终于在一件极端骇人听闻的事件中，姑娘被活生生吓死。这时候，大楼变得灯火通明，病人脱掉伪装，取走姑娘身上的电子背心，高价卖给神秘的买家。原来电子背心中记录了一颗健康的心脏，在高兴、害怕、惊恐，以至于骤然停止跳动的全部数据。这可能是笔者读过的第一篇恐怖小说，至今仍记忆犹新。

现在人们获取医疗数据，却变得相当简单。只要在手腕上佩戴一块类似电子表的仪器，就能随时随地把脉搏、体温、血压等数据，源源不断地传输到医疗中心。这些数据除了可以检测人们的健康以外，更是医疗保险公司的最爱。保险公司的精算师，根据这些数据可以开发新的保险产品，或者优化他们现有的产品组合。

从上面各种事例中，可以得出以下结论：第一，数据是无价之宝；第二，价值虽有，但确如沙滩中的黄金；第三，数据融合的价值，要远远大于种类单一的数据价值。

在研究各行各业数据应用时，笔者发现很多公司坐拥金山，却是苦苦挣扎。他们没有认识到自身的数据中正蕴涵着业务的重生之道。最早重视数据价值的是互联网公司，在大数据研究和应用方面领风气之先。但是，大数据并非仅仅是大公司的专利，它更多的是看待世界、产业的观念和视角。大公司自然可以合纵连横，跨界扩张；小公司也可以静水流深，别具高格。关键是你怎么看。

多快才算快？

答案是小于 1 秒，客户的体验就在分秒之间。

这一条是传统的数据应用和大数据应用最重要的区别。过去的十几年间，金融、

电信等行业都经历了核心应用系统从散落在各地市到逐步统一到总部的过程。大量数据集中后，带来的第一个问题就是大大延长了各类报表生成时间。业界一度质疑，快速地在海量数据中提取信息，是否可行？

谷歌公司在这方面的贡献，无疑是开创性的。它的搜索服务，等于向信息业界宣布，1 秒钟之内就能检索全世界的网页，而且可以找到你想要的结果。在写作本段的时候，当用谷歌搜索关键词“大数据”时，提示“找到约 46 300 000 条结果（用时 0.37 秒）”。谷歌等于为大数据应用确立了一个标杆。如果超过 1 秒钟的数据应用，就会给用户带来不良的使用体验。甚至在某些情况下，如果应用速度达不到“秒”级，其商业价值就会大打折扣。下面来看一个营销的例子。

价格越贵的东西，人们购买时就会越犹豫，反复掂量自己的钱包。相反，价格越便宜的东西，人们购买时更多根据一时的喜好，呈现冲动型购买的特征。京东商城根据消费者购买商品的特征，将其分为四种类型，其中冲动型购买者占 37%。冲动嘛，自然一闪即逝。所以能否在用户冲动的瞬间及时送达精准的商品信息，就成为了提高商品销售的关键所在。幸运的是，社交型互联网的应用，如美国的 Facebook、中国的微博和微信，提供了侦测人们偏好和兴趣的接口，使得这种精准的营销在大数据时代成为可能。

在以高频交易为主的股票市场，比别人快 0.02 秒，就可能获得惊人的超额收益。所以，有人为了抢这宝贵的 20 毫秒，单独建了一条从西海岸到东海岸横跨美国的光纤，也有人干脆就呆在纽交所相同的街区。这种毫秒级时差造成的商业机会，也许会随着大数据的普及应用而在其他行业不断上演。

以应急[①]为代表的一些新兴产业，对时效性要求非常高。假如市区某工厂发生事

① 应急产业一般指为预防、处置突发事件提供产品和服务而形成的活动的集合。按类别划分，一是救援处置装备与技术，二是监测预警诊断设备与技术，三是预防防护产品与技术，四是应急教育培训咨询服务等。应急产业具有多行业交叉和服务公共安全的属性，是新兴产业。发展应急产业，有利于国家的防灾减灾和公共安全，有利于基层的产业结构优化和社会和谐稳定，有利于企业的市场拓展和利润增长，有利于公众的安全和健康。

故，需在第一时间做出正确判断，第一时间评估影响范围，第一时间到达现场，第一时间开展正确的处置方法。

O2O[①]应用是互联网投资创业的一个热点领域。当消费者在商家门口经过时，就能收到商家的促销信息，这种服务听起来非常美妙。如果促销信息恰好是大家需要的商品或者服务，那么所有人都能从中受益。消费者节省了时间，商家卖出了商品，服务商获得了佣金。但是，如果推荐的不是消费者需要的商品，或者等消费者离开了才收到提示，就变成了令人烦恼的垃圾信息。没有人喜欢随时随地接收垃圾信息，垃圾信息和有价值的及时提示只有短短的几秒钟的差别。

再举一个信用卡消费提醒的例子。当笔者刷卡消费的同时，收到银行的提示短信，会感到很安全，也不会认为被打扰，因为当时正在处理跟消费支付相关的事情。如果过后才收到相同内容的短信，情况就不同了，也许笔者正在跟朋友聊天，也许正在写一篇文章，这条短信就成了打扰笔者的垃圾信息。客户的体验就在这短短的分秒之间。

孤立的数据是没有价值的

Facebook、微博为代表的社交网络应用，构建了普遍关联用户行为数据。本来大家在网络上浏览网页、购买商品，游戏休闲等等，都是互不关联的。尤其是智能手机的普及，大家的网络行为更趋向于碎片化。这些碎片化数据如果没有关联，是难以进行分析并加以利用的。但是社交网络提供了统一的接口，让大家无论是玩游戏还是买商品，都能够方便轻松地分享到微博上。微博扮演了用户行为数据连接器的角色。用户在网络上的碎片化行为，经由社交网络，就能完整地勾勒出一幅生

① O2O 即 Online To Offline，也即将线下商务的机会与互联网结合在一起，让互联网成为线下交易的前台。这样线下服务就可以用线上来揽客，消费者可以用线上来筛选服务，还有成交也可以在线结算，很快做到规模化。

动的网络生活图景，真实地反映了用户的偏好、性格、态度等等特征，这其中蕴育了大量的商业机会。

反之，孤立的数据，其价值要远远小于广泛连接的数据。然而，数据孤岛现象普遍存在。个人计算机中的文件，虽然按照目录分门别类的存放，但是之间的内容关系往往杂乱无章。企业中各部门壁垒林立，大家更倾向于尽可能地保护自己的数据。我国政府部门的数据孤岛现象更为严重，甚至可以称为“数据割据[①]”现象。在数据孤岛的影响下，难以发挥大数据中蕴藏的价值。

所以，笔者曾经和一些专家、学者交流，提到培育大数据能力的三个发展阶段。第一阶段，融合结构化和非结构化数据，消除数据孤岛现象；第二阶段，融合企业内部和外部的数据，消除数据割据现象；第三阶段，建立数据驱动的新型企业。对这三个阶段的探讨超出了本章的范围，后续还将有详细的描述。

活性越高价值越大

有一家公司给笔者寄来数据样本，希望帮他们评估这些数据的潜在商业价值。虽然数据量很大，但是数据更新的频率大概是每月一次。这样的数据类型很常见，一些支付公司收集的没有交费记录就属于这种情况。

所谓活性，也就是数据更新的频率。更新的频率越高，数据的活性越大；更新的频率越低，数据的活性越小。一般而言，数据活性更高的数据集，蕴含更丰富的信息。所以，这家公司如果想在大数据领域有所作为的话需要想办法提高数据的活性。

在判断公司的投资价值时，笔者挂在嘴边的一句话就是，要看公司拥有数据的规模和数据活性。之所以没有提多样化、快速等特征，是因为这样一句简练的话，更容易被大家理解和记忆。

---

① “数据割据”、“数据孤岛”是数据治理中最突出的两类问题。

第三节 大数据的认知框架

提要：

1.“三大发展趋势，六种商业模式”是本书解读大数据的认知框架。

2.“数据成为资产”是最核心的产业趋势，以数据资产为核心演绎出租售数据、租售信息、数据使能、数字媒体、数据空间运营和大数据技术提供商六类商业模式。

3.围绕数据资产，产业间拉开融合、分立的大幕。具体到信息产业内部，表现为靠近最终用户的公司在产业链上拥有越来越大的发言权。携用户优势，具备向产业链上游逆向整合的潜力。同时，产业链上游企业则积极向下游拓展。整体上呈现垂直整合趋势。

4.泛互联网化是积累数据资产，形成竞争壁垒的重要范式。大型公司如苹果、谷歌、亚马逊都是泛互联范式的典型公司。这也是有小型公司发展壮大的契机和路径。

资本市场观察大数据的态度是中立的，最基本的出发点是要识别哪些是真正创造价值的公司，而哪些又是“挂羊头，卖狗肉”骗股东、股民钱的“坏人”。所以必须深入到细节、必须洞察未来趋势、必须提出自己完整的理论和模型，不能人云亦云。说白了就是“练好一双火眼金睛，给妖精们当头一棒，让取经人拿到真经”。

在 2011 年 9 月，笔者注意到业界在大数据领域的发展动向后，随即开始了系统的调研分析，先后走访了 IBM、甲骨文、EMC、微软等行业巨擘，和国内 A 股上市公司、领风气之先的互联网公司、各大咨询机构、高校、研究所充分交流，连续发布了三篇大数据专题研究报告，持续跟踪海内外大数据领域的进展，逐步形成了相对完整的认知框架，如图 1-12 所示。

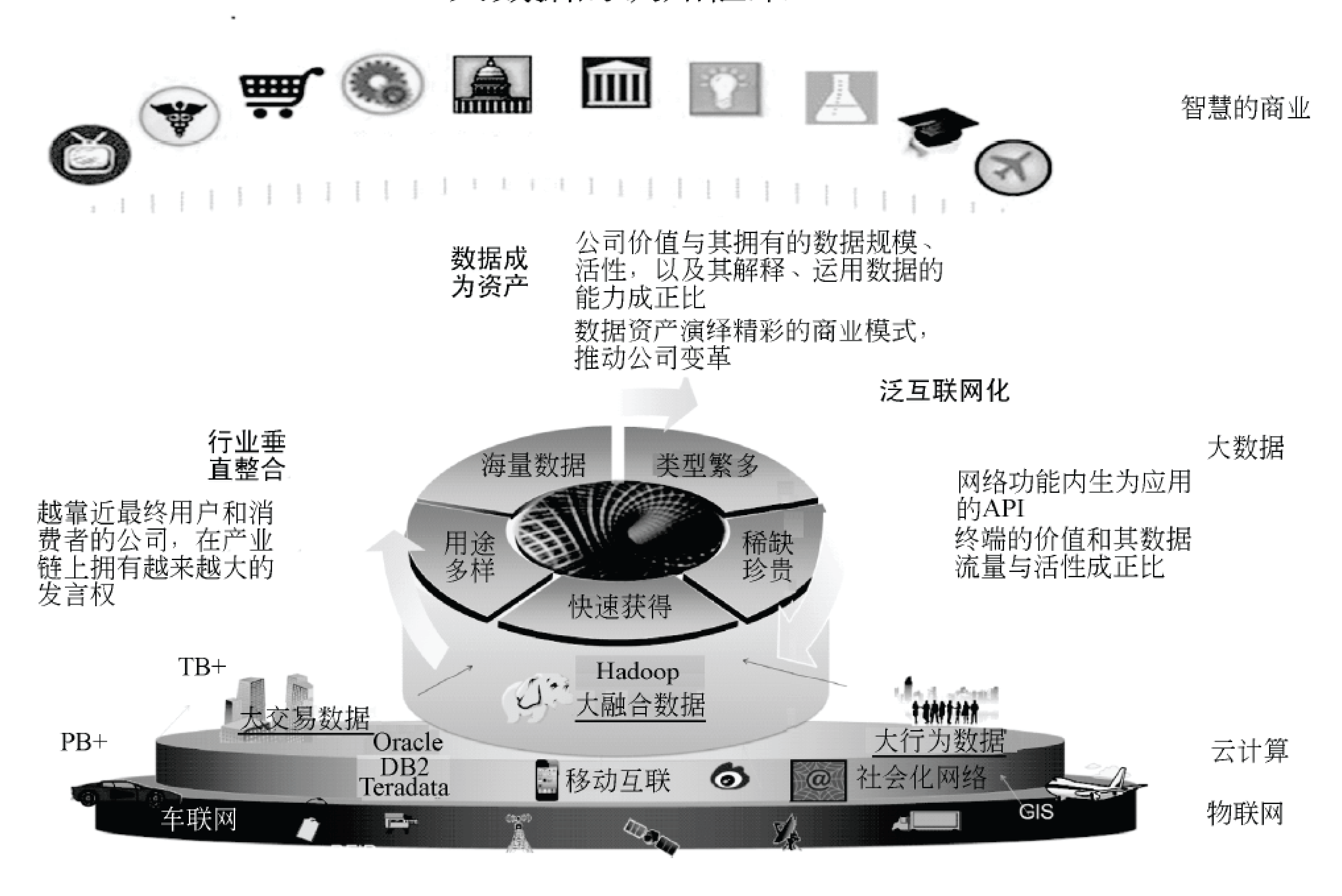


图 1-12　大数据认知框架①

围绕数据和最终用户，笔者观察到信息产业的发展具有三大趋势：第一，数据成为资产；第二，行业垂直整合；第三，泛互联网化。数据成为资产，更强调数据的战略意义；行业垂直整合趋势在数据运用层面，通过搜集大量的用户数据，更贴近用户、更理解用户，为其提供更适当的服务；泛互联网化驱动大数据飞轮效应的第一步，是收集数据的重要渠道，没有泛互联网化的应用软件和硬件设备，公司就难以获得用户的行为数据。三大趋势的提出，拓展了大数据主题的研究范围，开辟了新的视角和逻辑来观察信息产业内公司的成长路径和投资价值，成为分析研究的顶层逻辑的要素之一。

数据成为资产

数据成为资产是本书的重点内容和华彩章节，第三、四、五章都与数据资产内

① 参见国金证券大数据系列研究报告第二篇《大数据时代的三大发展趋势和投资方向》，第5页。

容相关。数据已经成为工业化与信息化深度融合的关键枢纽，成为推动产业融合兼并的战略资产，成为各地方城市转换发展思路的新思维，成为推动公司跨行业转型的根据地，成为数学与工程实践结合的最佳演练场。

在信息时代，数据将成为独立的生产要素。有人把“数据”比喻为工业时代的石油，但笔者认为“数据”和农耕时代“土地”的属性更加接近。如果企业拥有某类相对完整、全面的数据，退可偏安一隅，进可跃马中原。

谷歌、Facebook、亚马逊这三家互联网巨头，积累了不同的数据资产。谷歌为全世界的公开网页建立了最为庞大的索引；Facebook 拥有的社交网络积累了全世界最为庞大的人际关系数据库；亚马逊网站上沉淀了大量的商品信息，成为互联网上最为庞大的商品数据库。不同的数据资产，决定了它们不同的战略选择和商业模式。在某种程度上，它们甚至取代了 IBM、微软等传统的老牌巨头，在引领产业的发展方向。

拥有独一无二的数据资产的公司，将会获得难以置信的发展速度，培育出令人叹为观止的商业模式。事实上，它们具备了颠覆、冲击其他行业的压倒性优势。除了上面提到的互联网巨头外，本书中还谈到了雅昌公司的案例。这家从传统印刷行业起步的公司，通过年复一年、日复一日的漫长积累，形成了人类历史上空前的“艺术品数据库”。凭借这些数据资产，雅昌涉足出版、展览、收藏、移动互联网、线下实体博物馆等多个行业，其未来亦难以估量。

行业垂直整合

第二大趋势是行业应用的垂直整合。如图 1-13 所示，新兴产业往往是以垂直整合的态势开疆拓土，但产品成熟后，产业链上专业分工则激发出惊人的创造力，并且成本也逐渐降低，优势逐渐转向水平分工格局。但是当下，信息产业中行业垂直整合趋势明显，是大数据效应改变产业竞争格局的一个缩影。了解这个趋势，可以解释很多公司的成长逻辑，真真是“三十年河东，三十年河西”。在这个趋势下，越

靠近终端用户的公司，在产业上拥有越大的发言权。微软的股价十年横盘，苹果的股价却一飞冲天，两大巨头之间的恩恩怨怨、此起彼伏是这个趋势最好的注脚。

过去大家买计算机，关注的是 CPU 主频、内存、操作系统等；现在入手 iPad，直观感受是酷不酷，没有人问 iPad 的 CPU 是几个核的。这标志着消费者的关注重点已经迁移到产品能否满足自身的个性化需求了。在企业级市场也一样有相同的趋势，不要讲你的数据库、主机又出了什么新功能，客户更多会问“你们能不能满足我业务的需要？”这个趋势的出现有两大原因：第一，通用的平台型软件逐渐同质化；第二，用户对自身业务的关注超过了对计算能力的追求。

其实，很多人都没有意识到软件同质化①的问题。笔者观察到，几乎每个大型的商业软件都有对应的开源软件，而且这些开源软件的功能和性能也已经可以满足大部分客户的需求了。在第六章表 6-1 中列出开源软件和商用软件的对比，以及开源软件的统计数据，此处不赘言。需要提醒的是，谷歌、Facebook 这种世界级的平台，其核心技术架构都是开源软件唱主角的。开源软件的兴起和繁荣客观上也加剧了软件的同质化。在这个趋势下，拥有大量的客户并了解客户业务需求的公司，将会迎来一波大的发展机遇。

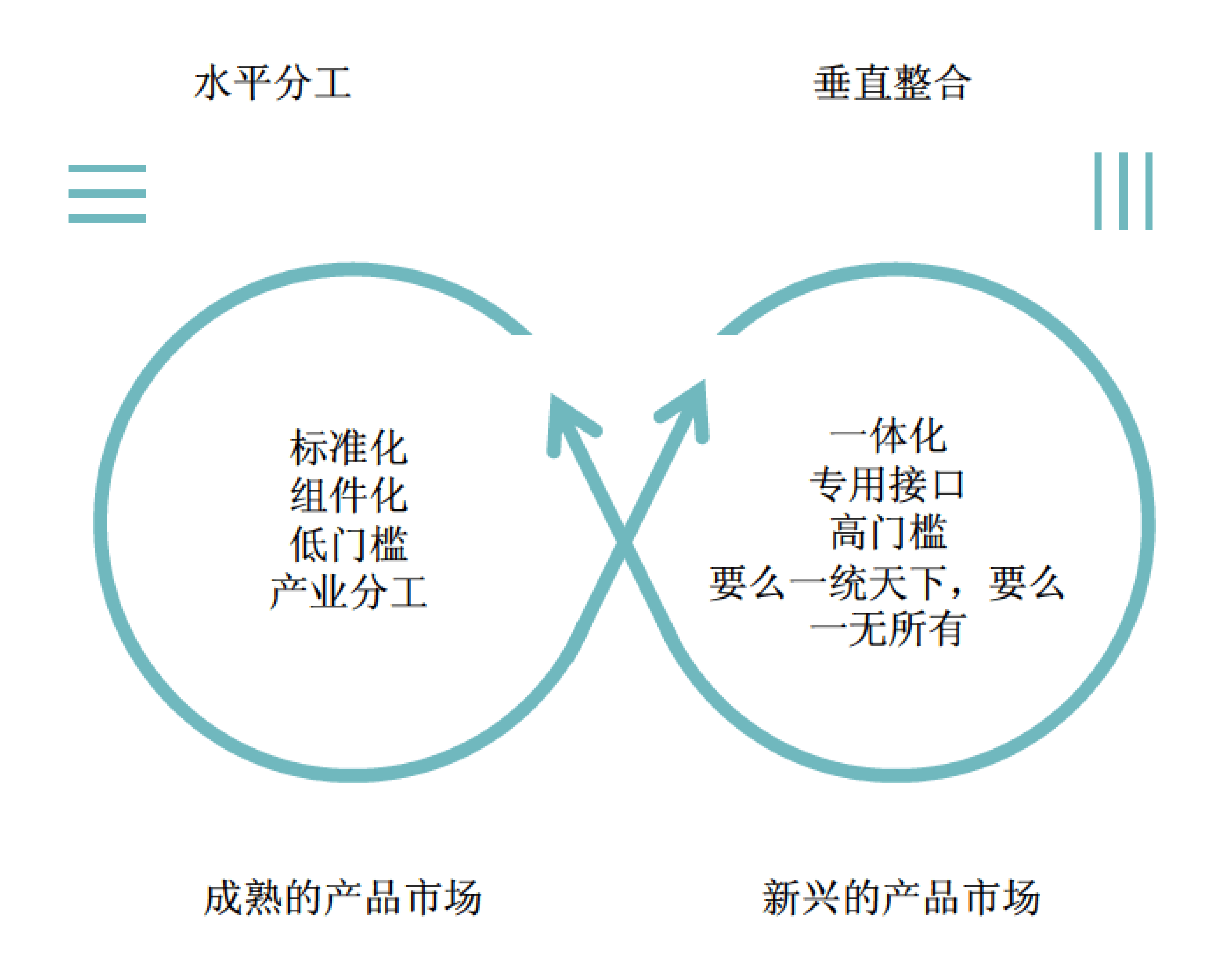


图 1-13　产业格局在垂直整合与水平分工之间摇摆

① 软件同质化，是从相对宏观的角度来审视基础软件的发展，更强调的是现在这个阶段用户的可替代选择增多，对单一厂商软件产品的依赖程度在不断的降低。

行业应用垂直整合的内容将在第六章展开论述。

泛互联网化

在讲述提出泛互联网化趋势时，提炼了一个重要思想——泛互联范式。在和产业界人士交流的过程中，笔者反复强调大数据并非只是大型公司的游戏，小公司、传统企业也一样可以搞得精彩纷呈。泛互联范式为其提供了现实可行的理论基础；亦是目前为止，实现大数据战略的最佳实践。

在泛互联范式中，强调终端、平台、应用“三位”加上大数据这“一体”，如图 1-14 所示。这四个方面都可以成为盈利的主要来源，但是，如果要取得竞争先机，则需要明确，主要靠哪部分盈利？需要补贴哪个方面？甚至在不同的发展阶段，盈利的主体也不尽相同。根据公司主要盈利来源的不同，可以简单归类成五种模式，分别是强终端模式、强应用模式、强平台模式、强数据模式以及混合模式。

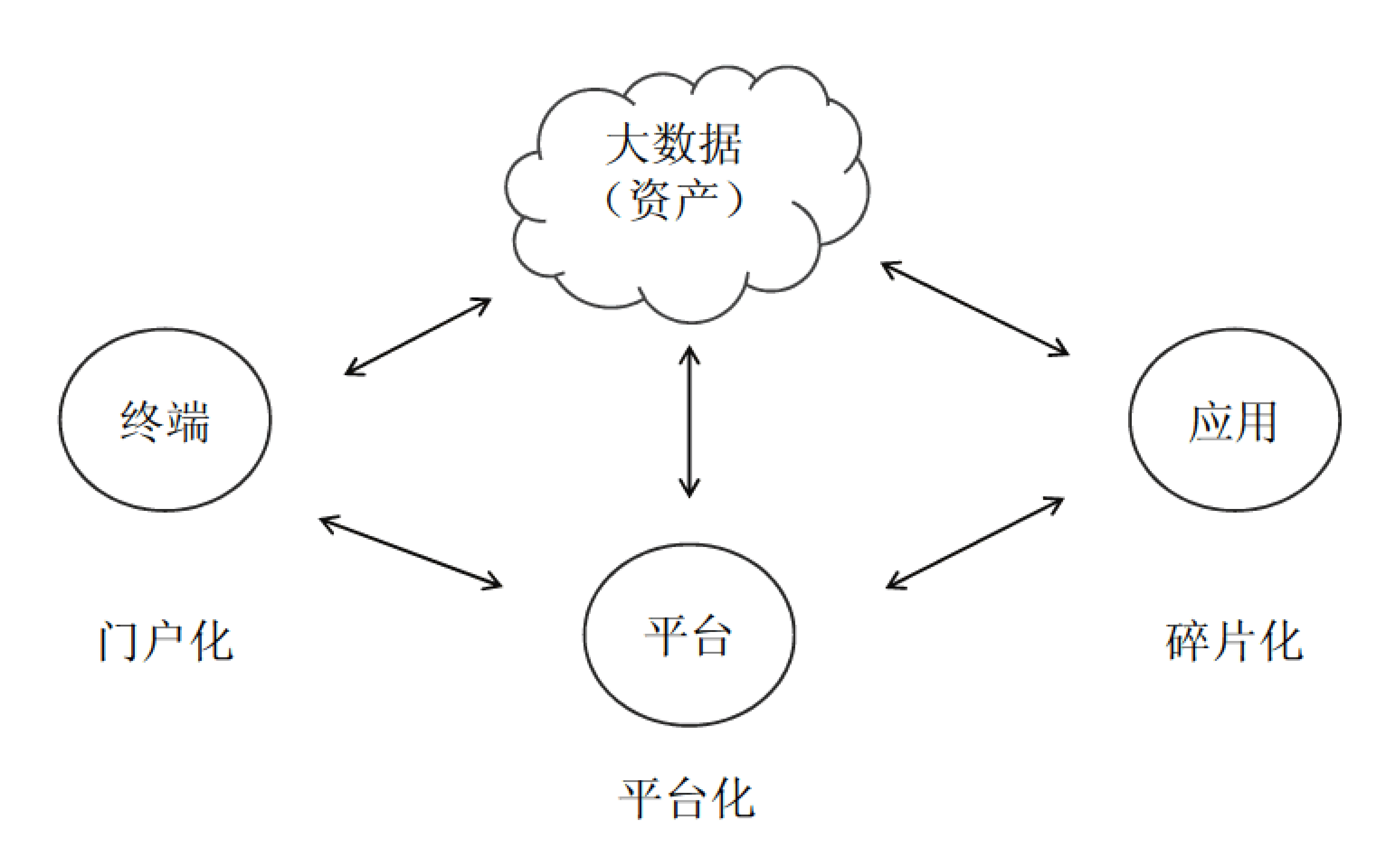


图 1-14 “三位一体”的泛互联范式

泛互联网化范式批判工业时代的标准化思维，指出利用科技手段，碎片化应用满足用户个性化需求才是王道。应用的碎片化，恰恰可以解决标准化产品和用户个性化服务间的矛盾。泛互联网化内涵非常丰富，以碎片化为例，事实上不仅仅应用

呈现碎片化趋势，服务、内容都可以碎片化适应新型媒介需求。譬如教育，如何满足人们利用零星时间学习知识的渴望呢？限于本书的篇幅，仅在第七章来阐释，先给出范式框架，再通过与各行各业的深度交流，不断地补充发展。本书第二版将会补充这部分内容。

提醒读者注意的是，传统企业如果灵活运用泛互联范式，往往能取得意料之外的高速增长。说一句很玄的话，“运用之妙，存乎一心。”

六种商业模式简述

围绕数据资产，考察不同行业的盈利方式和经营策略，归纳总结了六种商业模式。

1. 租售数据模式：简单来说，就是出售或者出租广泛收集、精心过滤、时效性强的数据。这也是数据成为资产的最经典的诠释。按照销售对象的不同，又分为两种类型：一是作为客户增值服务，譬如销售导航仪的公司，同时为客户提供即时交通信息服务。广联达公司为它的客户提供包年的建筑材料价格数据，仅此一项业务，年收入超过 1 亿元人民币。二是把客户数据有偿提供给第三方，典型的如证券交易所，把股票交易行情数据授权给一些做行情软件的公司。

2. 租售信息模式：一般聚焦某个行业，广泛收集相关数据，深度整合萃取信息，以庞大的数据中心加上专用传播渠道也可成为一方霸主。信息指的是经过加工处理承载一定行业特征的数据集合。

3. 数字媒体模式：这个模式最性感，因为全球广告市场空间是 5000 亿美元，具备培育千亿级公司的土壤和成长空间。这类公司的核心资源是获得的实时、海量、有效的数据，立身之本是大数据分析技术，盈利来源多是精准营销和信息聚合服务。

4. 数据使能模式：如果没有大量的数据，缺乏有效的数据分析技术，这些公司的业务其实难以开展。譬如，阿里金融为代表的小额信贷公司，通过在线分析小微企业的交易数据、财务数据，可以解决应提供多少贷款、多长时间可以收回等关键问题，把坏账风险降到最低。

5. 数据空间运营模式：从历史上来看，传统的 IDC 就是这种模式，互联网巨头都在提供此类服务，但近期网盘势头强劲。从大数据角度来看，各家纷纷嗅到大数据的商机，开始抢占个人、企业的数据资源，海外的 Dropbox、国内的微盘都是此类公司的代表。这类公司的发展空间在于可以成长为数据聚合平台，盈利模式将趋于多元化。

6. 大数据技术提供商：从数据量上来看，非结构化数据是结构化数据的 5 倍以上，任何种类的非结构化数据处理都可以重现现有结构化数据的辉煌。语音数据处理领域、视频数据处理领域、语义识别领域、图像数据处理领域都可能出现大型的高速成长的公司。

第四节 数据科学——改变探索世界的方法

提要：

1. 越来越多的事物不断的数据化，使得人们可以从大量的数据中，发现隐藏的自然规律、社会规律和经济规律。从这个角度来看，大数据将拓展人类的视野。
2. 大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑战。它将对现有的科研和教学体制带来大幅度的变革，对科学与产业之间的关系、科学与社会之间的关系带来大幅度的变革。

深入思考大数据带来的颠覆性的影响，其根源就是越来越多的事物数据化了。图像、声音、人类的情绪和基因组，看起来风牛马不相及，但是信息科技的发展都把它们神奇地变成了“0”、“1”的不同组合，也就是“数据”。

当网页变成数据后，谷歌具备了令人艳羡的全文搜索能力，在几毫秒之内，就

能让人们检索世界上几乎所有的网页。当方位变成数据后，每个人都能借助 GPS 快速到达目的地。当情绪变成数据后，数据科学家们甚至可以根据大家快乐与否判断股市的涨跌。这些不同的数据可以归结为几类相似的数学模型，从而使得“数据科学”成为一门具备普遍适用性的学科。譬如生物信息学、计算社会学、天体信息学、金融学、经济学、电子工程等学科，都依赖数据科学的发展。

事实上，数据科学还带给人们观察世界的新方法——从大量的数据中，揭示世界运行的规律。2008 年，《连线》杂志主编克里斯 • 安德森[1]就指出“数据爆炸使所有的科学研究方法都落伍了”，用一系列的因果关系来验证各种假设和猜想的研究范式已经不实用了，如今已经被无需理论指导的纯粹的相关关系研究所取代。安德森指出：“现在已经是一个有海量数据的时代，应用数据已经取代了其他的所有学科工具。而且只要数据足够多，就能说明问题。如果你有一拍字节的数据，只要掌握了这些数据之间的相关关系，一切就都迎刃而解。”[2]

安德森的观点引起轩然大波，但是的确值得深入思考。从牛顿力学到量子力学，科学家们建构了精巧的模型，原则上来讲几乎可以解释日常所有的自然现象。量子力学提供了研究化学、材料科学、工程科学、生命科学等几乎所有自然和工程学科的基本原理。但是保罗 · 狄拉克[3]指出，如果以量子力学的基本原理为出发点去解决这些问题，那么其中的数学问题太困难了。如果人们利用更为简单的数学模型，利用大量的数据则可以得到在工程实践中完全可行的结果。

人们在研究自然语言处理方面走过的弯路，为安德森的观点提供了有利的证据。20 世纪 50 年代，几乎所有的科学家都认为如果让计算机来充当翻译，就必须像人一样，让它理解词句的含义。于是提出人工智能的概念，让计算机来学习人类的各

---

① 克里斯 • 安德森（Chris Anderson），美国《连线》杂志主编，喜欢从数字中发现趋势。他是经济学中长尾理论的发明者和阐述者。著有《长尾理论》(《The Long Tail》)、《免费：商业的未来》(《Free:The Future of a Radical Price》)。

② 参见《大数据时代》([英]维克托 • 迈尔 • 舍恩伯格 肯尼思 • 库克耶著）第 92 页。

③ 保罗 • 狄拉克（1902—1984）全名 Paul Adrie Maurice Dirac，英国理论物理学家，量子力学的创始者之一。

种规则。这种方法很快在 70 年代走到了尽头。但是基于大量数据、运用概率模型的统计语言学的出现使得自然语言处理柳暗花明。如果没有这些概率统计模型，风靡一时的 Siri（个人语音处理）等应用，就不可能实现。

本书第九章将系统地阐述大数据给科学和教育事业提供的前所未有的机会，并指出：第一，数据科学将成为科研体系中的重要部分，并逐渐达到与包括物理、化学、生命科学等学科在内的自然科学分庭抗礼的地位；第二，数据科学研究和市场、产业有着密切的联系，在数据科学领域，从科学原理的发现到产业化所花费的时间远远短于传统科学的领域；第三，数据科学同样和人们的日常生活紧密关联。

第五节 大数据面临的挑战和机遇

提要：

1. 大数据将强烈冲击人们的观念，旧有僵化思维将导致公司在竞争中落了后手。片面地、孤立地、静止地看待大数据都是缺少大数据思维的典型特征。
2. 大数据时代“自主版权”将退居次席，信息技术本身的重要性将让位于数据资产的重要性。
3. 数据治理必须提到重要的地位，宏观层面国家解决“数据割据”问题需要顶层设计，企业则需要在“数据孤岛”间架起桥梁；微观层面则需要注重“数据质量”，包括数据的正确性、完整性、一致性。
4. 目前缺少必要的法律法规界定数据资产的归属和使用，客观上存在发挥数据资产商业价值与侵犯个人隐私之间的矛盾；缺少大数据人才，缺少系统的学科建设亦是制约大数据发展的关键因素。

大数据概念刚刚提出，有人击节赞叹，认为“数据人”的春天到了，也有人质

疑为炒作，认为不过是业界和资本市场又一次发神经而已；但更多的人是茫然的，并不知道这个概念对自己的业务意味着什么。本节主要澄清一些概念和误读，探讨大数据落地存在的障碍。

重新审视“自主版权”

大数据时代，产业重心发生了迁移。信息产业的重心由基础软件向应用软件过渡，信息技术本身的重要性向数据资产的重要性过渡。而应用软件领域，恰恰是中国软件企业的强项。利用好开源的基础软件，实现在应用软件领域的突破，带动基础软件领域的进步，是中国信息产业的发展方向。

“智慧出，有大伪”。有多少人假“自主版权”之名，却从未超越开源软件的功能？信息产业的创新，是亦步亦趋么？微软有操作系统，我们就必须搞“自主版权”的操作系统？多年的拨款，支持“创新”，为我国信息产业技术提升带来哪些进步呢？幸而我们有一个华为，看看华为老板任正非怎么说。

2012 年 7 月，任正非与华为实验室的干部和专家座谈。有人问：“当前在终端 OS 领域，Android、iOS、Windows Phone 8 三足鼎立，形成了各自的生态圈，留给其他终端 OS 的机会窗已经很小，请问公司对终端操作系统有何期望和要求？”

“如果说这三个操作系统都给华为一个平等权利，那我们的操作系统是不需要的。为什么不可以用别人的优势呢?微软的总裁、思科的 CEO 和我聊天的时候，他们都说害怕华为站起来，举起世界的旗帜反垄断。我给他们说我才不反垄断，我左手打着微软的伞，右手打着 CISCO 的伞，你们卖高价，我只要卖低一点，也能赚大把的钱。我为什么一定要把伞拿掉，让太阳晒在我脑袋上，脑袋上流着汗，把地上的小草都滋润起来，小草用低价格和我竞争，打得我头破血流。我们现在做终端操作系统是出于战略的考虑，如果他们突然断了我们的粮食，Android 系统不给我用了，Windows Phone 8 系统也不给我用了，我们是不是就傻了？同样的，我们在做高端芯片的时候，我并没有反对你们买美国的高端芯片。我认为你们要尽可能

地用他们的高端芯片，好好地理解它。只有他们不卖给我们的时候，我们的东西稍微差一点，也要凑合能用上去。我们不能有狭隘的自豪感，这种自豪感会害死我们。我们的目的就是要赚钱，是要拿下上甘岭。拿不下上甘岭，拿下华尔街也行。我们不要狭隘，我们做操作系统，和做高端芯片是一样的道理。主要是让别人允许我们用，而不是断了我们的粮食。断了我们粮食的时候，备份系统要能用得上。”

在国家“信息安全”的背景下，我们的确是要搞操作系统，万一别人不给我们用了呢？不能被人卡脖子。这是国家或者和华为一样体量的公司，不得不在安全层面思考的一个问题。但是过分强调“自主版权”的操作系统是否是任正非口中“狭隘的自豪感”呢？

国家的数据安全，应该建立在“自主可控”的软件、硬件之上，并非一定是“自主版权”的软件、硬件。自主可控与自主版权仅仅两字之差，但导致的产业方向截然不同。

华为过去没有自己的操作系统，也没有自己的芯片，但是硬是在广阔的“应用市场”打开一片天地。利用“应用”带来的市场地位、积累的研发实力，开始向产业链上游扩张。这是一条实实在在的路。华为的成功和战略选择，带给信息产业宝贵的经验，就是扎扎实实做好应用，切切实实积累技术。华为并不是在平地起高楼，充分利用了“开源软件”是华为在基础软件领域快速赶上的原因之一。在开源的Hadoop（大数据主流技术）社区重要贡献公司名单中华为排名第七，是贡献最大的中国公司。

过分的强调“自主版权”，使一些“头脑灵活”的公司嗅到“商机”。去开源软件社区，下载几个软件，改改界面，换一个标识，就成了“自主版权”软件，拿来骗取国家的科技补贴。这样的公司就是国家的蛀虫，产业中的败类。第一，欺骗国家；第二，违背开源社区的精神。这些公司的出发点从不是着眼于实际的应用，他们只是骗取国家的创新扶持拨款。他们的技术从开源社区“偷窃”而来，从无超越开源软件的可能。

相反，哪些埋头解决客户的实际业务问题，利用开源软件弥补自身基础软件的短板，在实际应用中不断地修改、完善、升华开源软件的公司，才是中国信息产业的希望，他们才有可能借助应用为王的时代实现反超。

充分利用开源软件，尊重开源社区分享、合作的精神，发展“自主可控”的基础软件、基础硬件产品，才是一条正路。事实上，中国绝大多数的软件公司都在利用开源软件。最值得学习和推崇的是华为公司。第一，他们大张旗鼓地在用，尊重开源精神；第二，他们不断地反哺开源社区，促进开源软件的发展。反哺开源软件是一种态度，更是一种能力。如果公司不能超越开源软件，是谈不上反哺开源的。除华为之外的第二类是偷偷地用，模糊版权问题，谈不上反哺开源社区。第三类则最为恶劣，明明是拿的人家开源软件，非要说自主版权，这种行径与偷盗无异。幸好中国有一些有志于技术的年轻人，自发地成立开源技术小组。笔者衷心地祝福他们在开源的道路上走得更远。

中国的互联网公司在使用开源软件方面做出了表率。淘宝网光棍节一天的销售额就达到了 191 亿，这在世界上都是独一无二的。这套以开源软件为基础构建、开发的后台信息系统可以说承受了最大的压力。京东商城也是如此，2012 年初，京东开始“去贵族化”（抛弃昂贵的商业软件）的努力，以开源软件为主，重新构建了其信息系统。笔者在和其 CTO 交流的时候，他说感到非常欣慰，因为这次光棍节的购物，京东的信息系统没有出现任何性能问题。海外最大的电子商务公司亚马逊、最大的搜索引擎谷歌、最大的社交网站 Facebook，无一例外都选择了以开源软件为主构建信息系统。而且大数据技术，本就是开源软件唱主角。既然如此复杂的业务、如此巨大的交易量都可以使用开源软件，我们为什么要花大把大把的金钱给那些提供昂贵产品的公司呢？京东商城恰好又是非常典型的例子：京东的 CTO 是从大名鼎鼎的甲骨文（Oracle）公司挖来的，但也正是他主导了京东“去甲骨文”的历程。

开源软件是送给中国信息产业界的一份大礼，我们要大大方方地接受它、改造

它、支持它。这是一种态度，更是一种能力。校正公司对待开源软件的态度，引导公司加强开源软件研发、改进，支持开源事业，则是信息产业政策需要认真对待的一个课题。开源软件既然是送给我国信息产业的一份大礼，那么如何收下、如何用好，就是需要政府和产业界共同面对的大命题了。

缺少大数据思维和意识，没有紧迫感

曾经有人问，发展大数据要采用哪些技术，有什么产品？事实上，大数据首先是一种思维方式，其次才是判断产业发展趋势和选择公司战略，最后才谈得上技术实现的问题。有四种典型的片面认识阻碍企业家完整地认知大数据：第一，认定是炒作；第二，片面理解；第三，视野狭隘；第四，唯技术论。这些都是缺少大数据意识的表现。尽管还有其他各种客观原因，但是企业家的思想认识是阻碍大数据获得深入应用的最重要因素。

第一，认定无非是另一次炒作。这是最常见的一种误读，其流毒在于阻碍了人们去耐心认真地研究大数据的由来和机理。IT 业和资本市场的确有炒作的传统。对“千年虫”连篇累牍的报导和宣传，除了让 IBM 等公司大赚一笔外，结果发现问题并没有事前描述的那么耸人听闻。物联网也曾经是资本市场的宠儿，但现在却已风光不在。如果因此就把大数据归于炒作一途，肯定会与机会失之交臂。大数据与以往的技术概念有显著的不同，最大的差异是大数据已经远远超越技术的概念，是互联网、智能终端、社交网络发展到一定阶段的必然产物。以往，信息技术总是在围绕提升企业运营效率打转，而大数据促使商业智能真正走向企业的决策中枢。

第二，片面的理解。有人一听说大数据，就说十多年前我们就有多少多少数据，以前都说海量数据如何如何。的确，海量数据是大数据的特征之一，但海量数据并不等同于大数据。大数据更强调数据的多样性、及时性。网络日志、文档、视频、图片等都是大数据关心和处理的对象。更重要的是，大数据技术总是要求尽可能快地发现有决策价值的信息，快的度量单位是不能超过 1 秒。厂商在介绍大数据概念

时往往介绍三个“V”特征：Volume，体量大，至少要到 PB 级别（1PB=1024 TB，大约相当于地球观测系统五年的数据）；Velocity，实时要求高；Variety，强调数据的多样性。还有厂商增加一个 V——Value，意思是说大数据有价值。这些都是对的，但不免过于片面。

第三，狭隘的视野。仅仅埋头在自己的一亩三分地，是难以领略大数据全部魅力的。首先它是超越行业的，一定会促使新的行业诞生，也一定会令一些行业消亡，几乎所有行业的竞争格局都将被大数据所颠覆；其次它是超越技术的，无论是开源的 Hadoop，还是各厂商力推的新产品，都不足以反映大数据的全貌。作为投资人或者公司的决策者，如果不能确立这是行业竞争的战略要地思维，则不足以妄谈大数据。

以企业在线服务市场为例，这个看起来很朝阳的产业，并没有在中国取得引人注目的成长。国内最大的几家公司，营业收入大约在 1 亿元左右。笔者曾和业内人士辩论能否免费为企业提供在线服务。大多数业界人士认为企业市场与个人市场不同，企业客户担心免费服务的质量，不收钱人家反而不敢用。事实上，笔者曾看到已经有公司免费为企业提供在线的企业管理服务，其盈利模式变成为它的在线客户提供金融贷款业务。在线业务加小额贷款服务已经成为极具颠覆性的商业模式，这种商业模式如果进展顺利，传统的在线服务商将面临行业性的灭顶之灾。这种新模式，其核心竞争力体现在拥有大量的、真实的客户运营数据，借助对这些数据的收集分析，预测客户的运营风险，最大限度地降低借贷违约风险。阿里巴巴公司刚刚提出的平台、数据、金融的战略，则是大数据前景的最佳诠释。

广告产业将重新洗牌。大家都知道广告预算至少有一半被浪费掉，可悲的是不知道浪费的是哪一半。借助大数据技术，广告将变得及时和精准，而且能够评估量化每个渠道的广告效果，看起来具有非常诱人的前景：广告主大大节约资金，消费者得以避免垃圾广告的骚扰。理论上，如果大数据技术得到充分运用，那么我们每个人将不会收到垃圾信息。在日常消费中，冲动型的购买决策越来越普遍。商家必

须在消费者最感兴趣的时候，及时触发刺激消费者的购买欲望。离开大数据的支持，这种精准的营销则难以实现。

制造业将重新定义核心竞争能力。在制造业发展的不同阶段，其核心竞争力是不同的。在发展初期，产品质量是非常重要的因素，就是能够做到人有我优，这个阶段的关键资源是拥有先进的生产设备。产品同质化后，对于渠道的掌握和控制成为生命线，关键资源是优质经销商队伍。当渠道成熟到一定的阶段，谁能掌控终端，谁将占据竞争优势，关键资源是终端营销团队。考察制造业关键资源的迁移，笔者发现它逐渐向最终用户端迁移。换句话说，谁能掌握最终用户，谁就能笑傲江湖。这方面的例子还有很多，各行各业都不在少数。对此本章不在赘言，后续章节均有详细描述。

第四，唯技术论。大数据是一种思考方式，和有没有数据、数据量的大小、使用什么技术，不存在严格的正相关。没有最新的技术，也可以通过数据资产来获利。即便拥有最先进的技术，缺少数据思维，没有数据资产，往往也徒劳无功。不能单纯地认为只有那些围绕 Hadoop（泛指大数据技术）开发的新兴公司，才是大数据公司，也不能认为没有技术的就不是大数据公司。相反，在大数据领域，那些拥有稀缺性数据资产的公司往往可以指点江山，独领风骚。大数据既不等于数据挖掘也不等于统计分析，更不等于人工智能。但是这些技术和算法都需要大数据的支持。使用同样的算法，如果利用全部的数据集，而非小样本量，甚至可得出截然不同的结论。这就是大数据的魅力。它可以在宏观尺度上把握潮流，也可以在微观颗粒上预测未来。

数据治理缺位

数据割据、数据孤岛和数据质量是典型的三大数据治理问题。

因为制度、地方主义、部门主义等人为因素造成数据分散的现象，称为“数据割据”；因为技术差距、历史遗留问题等形成的数据分散的现象，称为“数据孤岛”。

数据割据现象更多存在于国家各部门、各地方之间，大型企业内部也会存在数据割据现象。

譬如气象部门详尽的天气观测数据，是研究大气规律、做天气预报的第一手资料。但是这些数据因为各种各样的原因在气象局那里睡大觉。理论上讲，科学院的大气物理研究所是可以拿到这些观测数据的，否则大气所的科学家们怎么支持气象局的工作啊？根据“有关部门的有关规定”，大气所的确也能够接触到这些数据。但实际操作中，要拿到这些有用的数据，不拖个半年是不行的，而且就算到手了，也是鸡零狗碎的，没什么用途。这就是典型的“数据割据”现象。

有家公司专门为淘宝网上的商家提供在线服务，但这些服务需要淘宝开放数据接口。早期，如果不使用淘宝提供的服务器是没有任何障碍的，但现在这项服务有50%的时间是无法连通的。我们自然无权指责淘宝的经营策略，但这种因先天优势进而形成数据割据的局面，的确令人担忧。

美国政府在消除数据割据方面可谓用心良苦，除了系统性地提出国家层面的数据战略外，一些做法也值得借鉴。具体方法参见本书第三部分的详细介绍。

我国政府面临更加严峻的数据割据困境。数据保护主义不过是部门保护主义在信息领域的延伸而已，必须出台国家级别的顶层设计，由上而下地破除阻碍数据分享的藩篱，并建立数据共享、成果分享的利益分配机制，才有望从根本改善数据割据的问题。

数据质量的好坏，直接影响数据资产的价值。数据质量主要包括数据的真实性、完整性、一致性。数据质量的解决非一日之功，需要技术、制度、文化等等方方面面的努力。如果把数据认认真真地当成资产对待，数据质量就是需要面对的第一个问题。

数据资产的界定与安全

随着数量越来越多的数据被数字化，数据在跨越组织边界而流动着，一系列政策问题将会变得越来越重要，这包括但不限于隐私、安全、知识产权和责任。显然，

随着海量数据的价值愈加明显，隐私是个重要等级（尤其是对消费者来说）不断提高的问题。个人数据（如健康和财务记录）经常能够提供最重要的人类福利，如帮助精准确定适当的医疗或者最恰当的金融产品。然而，消费者也将这些类别的数据视为最敏感的个人隐私。显然，个人和其生活所在的社会将不得不努力在数据隐私和数据的功用之间权衡取舍。

另一个密切相关的担忧是数据安全，例如，如何保护竞争方面的敏感数据，或应保持隐私的其他数据。最近的例子表明，数据被盗不仅暴露消费者个人信息和企业保密信息，甚至还会暴露国家安全秘密。鉴于严重的数据被盗事件有增无减，通过技术和政策工具解决数据安全问题将成为关键。

海量数据日益提升的经济意义也昭示了一系列法律问题，即数据与许多其他资产具有根本性的差异，尤其是当其与如下事实联系起来时，数据可以与其他资产结合起来完美而轻松地复制，同样一份数据可以由多个人同时使用。这些是数据与实体资产相比独有的特征。有关数据所附带的知识产权问题不容回避：何人“拥有”某份数据？某一数据集附带着何种权利？数据的“公平使用”的定义是什么？此外，还有与责任相关的问题：当一份不准确的数据导致负面结果时谁应负责？要充分发挥海量数据的潜力，此类法律问题需要澄清，这也许会随着时间的推移逐步澄清。

缺乏大数据人才

就算政府和企业界认识到大数据可以释放经济的下一波增长潜力，认识到数据资产是关乎企业未来的命脉。但是如果想要成功运用大数据技术，达成企业战略目标，最大的制约因素往往是大数据人才的匮乏。这一点已然成为推广利用大数据技术的阿喀琉斯之踵。

不过许多高校近期的举动令人欣慰。北京大学、上海交通大学、中国人民大学、北京航空航天大学等高校都在设立数据科学的专门研究机构和相关专业，未来，也许数据科学家将成为令人尊重的职业。

第一部分

产业大势

大势汤汤，顺昌逆亡。产业兴衰的决定性因素，已经不是一城一地的争夺。土地、人力、技术、资本这些传统的生产要素，甚至需要追随“数据资产”重新进行优化配置。数据成为推动行业融合兼并、企业做大做强的战略性资产。不同产业围绕“数据资产”展开的争夺，将重新定义产业的生态环境和竞争格局！

导读：

1. 从 2011 年到 2012 年，有五件事情是大数据时代已经到来的标志。
2. 2011 年 5 月，麦肯锡发布《大数据：创新、竞争和生产力的下一个前沿领域》报告，是国内外产业界的先声。
3. 2011 年 12 月，中国资本市场开始发布大数据系列报告，包括《大数据时代即将到来》、《大数据时代三大发展趋势和投资方向》、《以数据资产为核心的商业模式》等，引起资本市场和产业界的高度关注。
4. 2012 年 3 月，美国奥巴马政府发布《大数据研究和发展计划》，引起各国震动，标志大数据上升为国家战略，体现国家意志。
5. 2012 年 4 月，全球首家大数据公司 Splunk 在纳斯达克上市，上市当天市值达到 30 亿美元。当时这家公司尚未盈利。
6. 2012 年 11 月，我国首届数据科学与信息产业大会召开，标志我国学术界、产业界、资本市场形成共识，共同推进大数据的发展和落地。

第二章

大数据时代已经到来

大数据时代已经到来，且正在引发一场革命！

——笔者

2011 年底，笔者发布第一篇大数据报告的时候，用的标题是《大数据时代即将到来》。没想到短短一年，大数据领域发生的变化波澜壮阔，令人心驰神往。大数据已经不仅仅局限于信息技术产业，而已事关国计民生、经济大势，政府、学术界、产业界、资本市场都在行动。在这里简单梳理一下 2011 年到 2012 年的几件大事，作为大数据时代的见证，如图 2-1 所示。



图 2-1　大数据时代到来的标志性事件

麦肯锡于 2011 年 5 月发布报告《大数据：创新、竞争和生产力的下一个前沿领域》，将大数据概念从技术圈引入企业界。国金证券率先将大数据概念引入中国资本市场，连续推出三篇报告，令资本市场沸腾。巧合的是，美国政府在笔者的大数据研究报告发布不久就推出了《大数据研究发展计划》，将大数据上升至国家战略层面，形成国家意志。接下来，Splunk 成为在美国成功上市的首家大数据公司，让“数据人”一时扬眉吐气，深感数据工作的春天到了。2012 年 11 月 17 日在北京大学召开的“数据科学与信息产业大会”上，宣告数据科学将在大数据时代焕发新生，标志学术界对大数据的重视达到了一个前所未有的新高度。

正如哈佛大学量化社会科学学院院长 Gary King 所说：“这是一种革命，我们

确实正在进行这场革命，庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来，没有哪个领域不会受到影响。”毫无疑问，上述的种种事件无不向世界传递一个信息：大数据时代已经到来！

第一节　国内外产业界的先声

最早将“大数据”概念带出技术圈的机构是国际知名的咨询公司麦肯锡。2011 年 5 月，麦肯锡发布《大数据：创新、竞争和生产力的下一个前沿领域》研究报告，报告指出全球数据正在呈爆炸式增长，数据已经渗透到每一个行业和业务职能领域，并成为重要的生产因素。大数据的使用将成为企业成长和竞争的关键，人们对大数据的运用将支撑新一波的生产力增长和消费者收益浪潮。

麦肯锡深入研究了美国医疗卫生、欧洲公共管理部门、美国零售业、全球制造业和个人地理信息五大领域，用具体量化的方式分析研究大数据所蕴含的巨大价值。大数据的合理有效利用，为美国医疗卫生行业每年创造价值逾 3000 亿美元，为欧洲公共管理部门每年创造价值 2500 亿欧元（约 3500 亿美元），为全球个人位置服务的服务商和最终用户分别创造至少 1000 亿美元的收入和 7000 亿美元的价值，帮助美国零售业获得 60%的净利润增长，帮助制造业在产品开发、组装方面降低成本 50%。

通过对五大领域的重点分析，该研究提出了五种可以广泛适用的利用“大数据”的方法：

1. 创造透明度，使利益相关者更容易及时获取大数据产生的巨大价值。

2. 启用实验来发现需求，呈现可变性，提高性能。数据驱动的组织在已有经验成果的基础上做出决定，这种方法的好处已经被证实。

3. 细分人群，采取灵活行动。随着技术的进步，可以接近实时地进行细分，并通过更精确的服务满足客户需求。

4. 使用自动化算法代替或辅助人类决策，基于大数据的深入分析可以大幅降低决策风险，提高决策水平。

5. 创新商业、产品和服务，大数据使各类企业拥有了改善和创新现有产品和服务的机会，甚至建立全新的商业模式。

此外，麦肯锡在报告中也指出了在挖据大数据潜能时所面临的各种挑战，包括隐私、安全、人才、技术等。

麦肯锡的报告充分肯定了大数据蕴藏的巨大价值，并试图帮助不同地域、不同部门的领导者及政策制定者了解如何利用大数据的潜在价值。整篇报告为大数据时代的蓬勃发展拉开了序幕，引用其中的一句话作为评价，即“本研究绝不代表在大数据方面最后的话语，相反，我们把它看作是一个开始”。

第二节　中国资本市场反应敏锐

宏源证券计算机行业研究团队非常敏锐，团队的主要成员都在信息产业界有丰富的工作经验。笔者虽然现在从事行业分析师工作，但一直没有离开信息产业，从参加工作算起已经在信息产业混迹了 15 年。当系统地审视大数据潜力的时候，发现它的影响力必将超越信息业，推动各行各业向更自动化、更智慧的方向发展。如何把这个概念清晰地、系统地传达给每一位投资人，成为笔者研究大数据的原始推动力。

从 2011 年年底到 2012 年上半年，笔者陆续推出了三篇关于大数据的系列分析报告，依次为《大数据时代即将到来》、《大数据时代的三大发展趋势及投资方向》和《以数据资产为核心的商业模式》，首次在中国资本市场系统、全面地阐述了大数据潜在的巨大社会意义和经济意义，开资本市场大数据之先河。第一篇报告宣告了大数据时代的到来，并详细阐述了大数据的内涵、特征，大数据引发的变革以及大数据伴生的大机遇等内容。第二篇报告分析了大数据时代的三大发展趋势：一是软

件应用泛互联网化，呈现平台化、门户化和碎片化的特征；二是行业应用垂直整合，越靠近最终用户的企业，将在产业链中拥有越大的发言权；三是数据成为资产，数据就是金钱，对数据的掌控导致了对市场的支配和巨大的经济回报。第三篇报告梳理了以数据资产为核心的六种商业模式，即租赁数据模式、租售信息模式、数字媒体模式、数据使能模式、数据空间运营模式和大数据技术提供商。三篇报告一脉相承，层层推进，全面系统地分析了大数据时代背景下社会和产业的巨大变革及其衍生的产业机遇。

本书的第一部分内容就是依据上述三篇报告提出的大数据认知框架不断丰富和发展而成的。三篇报告的摘要，参见附录二。

第三节　美国政府的手笔

提要：

1. 美国奥巴马政府对信息科技的重视，超过历史上任一时期。数据改变了美国，是驱动美国转型的力量。
2. 美国的“大数据研究与发展”计划与 20 年前的“信息高速公路”计划一脉相承。
3. 美国发展大数据的思路是“众人拾柴火焰高”，国防部、能源部、卫生研究院、国家科学基金、地质勘探局、NASA 等部门提出尖端项目，产业界、学术界、非营利性组织、资本市场通力合作，共同推动大数据的发展与深化。

美国是信息革命的领军人

美国是最早预见和推动信息革命的国家。早在 1946 年，美国军方就研制出了

世界上第一台电子计算机——电子数字积分计算机，并在 1969 年由美国国防部高级研究计划局建立了计算机互联网 Internet 的最早雏形——ARPAnet。随后，在美国总统克林顿和副总统阿尔·戈尔的大力推动下，美国政府于 1993 年 9 月推出“国家信息基础设施(National Information Infrastructure，NII，亦称“信息高速公路”)”计划，其核心内容由建设覆盖全美的宽带高速信息网、利用信息资源研发传输编码与网络标准、开发制造信息设备、培养信息人才等几部分构成，旨在将美国所有的政府部门、公司、医院、学校、图书馆等不同机构以及每个家庭连接起来，建立一个计算机化的全国性高速信息网络。

美国“信息高速公路”计划开启了互联网时代，使人们的沟通、工作和生活方式发生了巨大的变革，同时也推动了信息技术和信息产业的快速发展，为美国赢得了新的经济增长点和长达十多年的快速发展期，成就了美国在互联网上的全球霸主地位。

美国沿“信息高速公路”，狂飙到“大数据”

美国奥巴马政府于 2012 年 3 月正式启动了“大数据研究和发展”计划，如图 2-2 所示，该计划涉及美国国防部、美国国防部高级研究计划局、美国能源部、美国国家卫生研究院、美国国家科学基金、美国地质勘探局六个联邦政府部门，宣布将投资 2 亿多美元，用以大力推进大数据的收集、访问、组织和开发利用等相关技术的发展，进而大幅提高从海量复杂的数据中提炼信息和获取知识的能力与水平。该计划并不是单单依靠政府，而是与产业界、学术界以及非营利组织一起，共同充分利用大数据所创造的机会。这也是继 1993 年 9 月美国政府启动“信息高速公路”计划后，国家层面发力在信息领域的又一次“狂飙猛进”。

随着网路技术、计算机技术以及通信技术的快速发展，人类社会的数据总量呈现指数级增长，根据谷歌前 CEO 现董事会主席埃里克·施密特的说法，截至 2003 年，人类社会总共创造了 5EB 的数据，而现在仅需要两天就能创造相同的数据量，

这对数据的收集、运输、存储、分析利用和安全等技术应用和数据的管理工作提出了更高的要求和更大的挑战。信息作为三大社会资源之一(另外两个是物质和能量)，如何充分利用信息资源，如何更经济更快速地从海量、不同结构类型的复杂数据中提取价值成为关键。

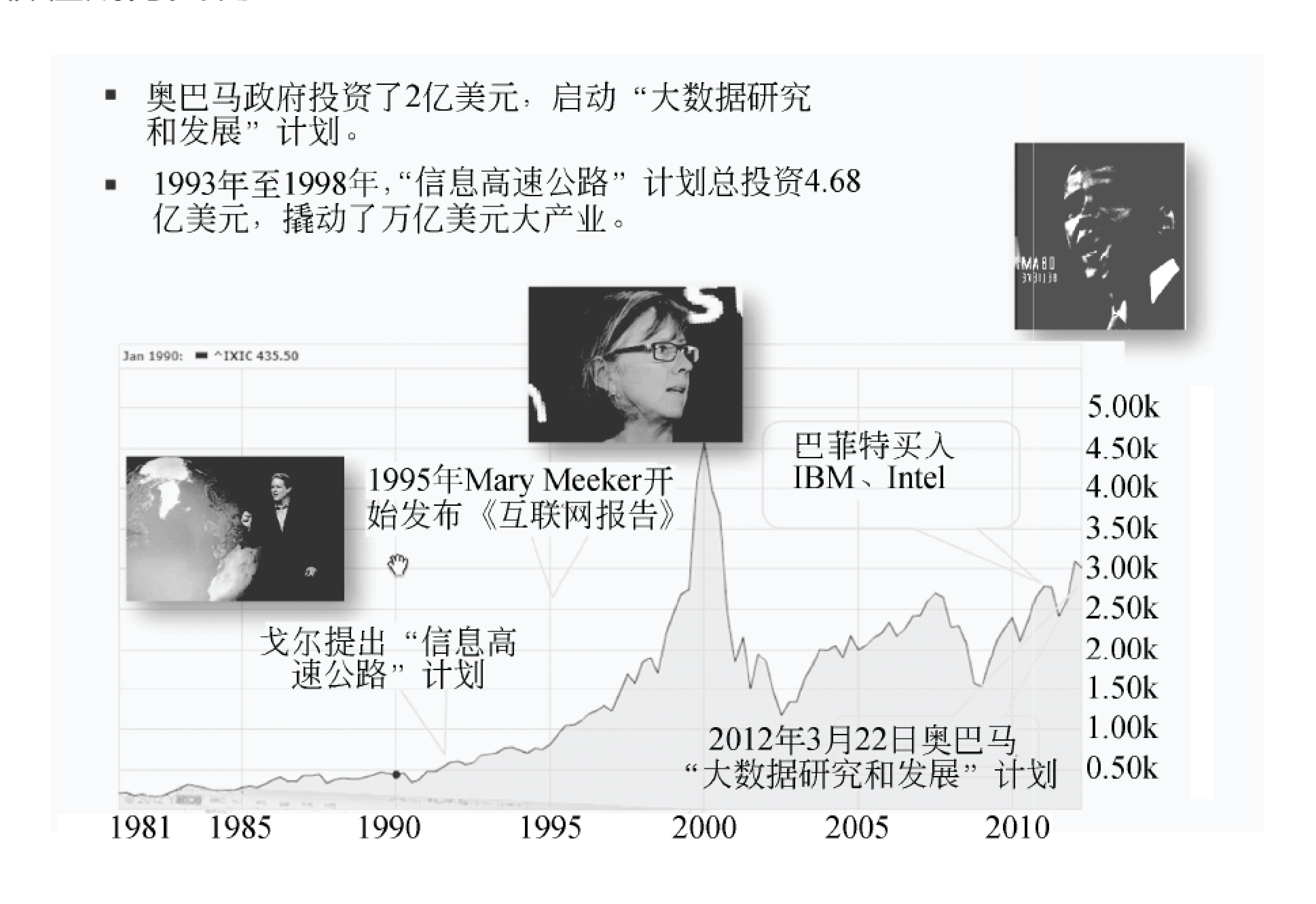


图 2-2　美国从“信息高速公路”计划过渡到“大数据研究和发展”计划

美国前总统克林顿开展的“信息高速公路”计划是通过高速率的通信网络搭建人们的信息交流网络，进而带动经济的快速发展。该计划促使海量数据的产生，但未能实现对数据资源进行充分利用，尤其是在大数据时代的今天，海量数据具有的巨大价值被白白浪费了。根据 2011 年美国总统科学技术顾问委员会提出的一份建议显示，大数据相关技术具有重要战略价值，但美国联邦政府对其研发投资却明显不足。而通过“大数据研究和发展”计划可以深度挖掘大数据的潜在巨大价值，带动产业的升级换代。从这个角度讲，“大数据研究和发展”计划与“信息高速公路”计划可谓一脉相承，层层推进。

美国将“大数据”上升至国家意志，影响深远

以奥巴马为首的美国政府推出“大数据研究和发展”计划，将大数据发展战略从商业行为上升到国家意志的层面，这将对未来十年科技与经济的发展带来深远的影响。

推动美国快速发展

“数”中自有黄金屋。大数据如果得到充分合理地开发利用，对个人、企业、政府乃至整个人类社会都具有极大的价值。美国推出的“大数据研究和发展”计划将对美国综合国力、国家安全、大数据技术、社会经济等方面产生巨大的影响。

增强国家综合国力。美国政府充分意识到了大数据的价值和意义，将其定义为“未来的新石油”，并指出一个国家拥有数据的规模、活性以及解释运用的能力将成为其综合国力的重要组成部分，对数据的占有和控制甚至会成为继陆权、海权、空权之外的另一种国家核心资产，成为世界各国之间的新博弈对象。由于以中国为代表的新兴国家的迅速崛起，美国在经济、政治等方面的影响力和控制力大幅下降，国际霸主地位面临巨大的挑战，“大数据研究和发展”计划有利于提高美国对数据资产的掌控能力，进而抢占新的国际战略制高点。

维护国家安全。信息网络突破了传统地域条件的束缚，在全球范围内实现了信息的快速高效流动，促进了人类社会的发展，但同时也带来了巨大的安全隐患。尤其是经济、政治、军事和科技等方面的核心机密信息的安全问题具有极高的战略意义，21 世纪国家信息安全的地位和作用不亚于传统的国防军事。如 2010 年 11 月，超过 25 万份由美国大使馆发出的机密电报被公开，泄露了美国在政治、军事战略等方面的机密信息，并引发了全球性的外交危机，这便暴露了国防部信息安全方面的诸多问题。美国国防部高级研究计划局在这次大数据研究方面给予了大量的投入，具体包括解决大规模数据集的异常检测和特征化的多尺度异常检测，开发新技术检

测军事计算机网络与网络间谍活动的网络内部威胁、加密数据的编程计算及视频图像的检索和分析工具等项目，这将大大提升美国应对大数据时代国家信息安全挑战的能力。

推动大数据技术发展。高速增长的海量复杂的大数据对传统技术提出新的挑战，研发大数据相关的核心技术成为当务之急。美国国家科学基金会以大数据关键技术的突破作为大数据项目的重点，与美国国家卫生研究院对大数据进行联合招标，包括管理、分析、可视化以及从大量多样数据集中提取有用信息等核心技术。此外，美国国家科学基金会还积极推进各项大数据相关措施，包括鼓励大学开发交叉学科课程，培养大数据人才，提供资金支持本科生使用复杂数据图形和可视化技术的培训等。上述的系列措施将有利于美国打造在大数据提取、分析等核心技术方面的竞争优势。

推广大数据应用，打造新的经济增长点。美国能源部推进高性能存储系统、千万亿次数据分析处理以及生物和环境研究计划，美国国家卫生研究院推进千人基因组计划数据的云端免费开放，美国地质勘探局推进地理系统科学的大数据，这些项目加快了大数据在能源、医学和地质等领域的应用及价值开发。据麦肯锡《大数据：创新、竞争和生产力的下一个前沿领域》研究报告显示，大数据为美国医疗服务业每年创造 3000 亿美元的经济价值。与此同时，在美国政府的推动下，企业、科研院校以及非营利机构将纷纷加入其中，进而形成利益相关者全体成员系统化共进的局面。目前，EMC、IBM、惠普、微软、Oracle 等 IT 巨头正积极通过并购实现技术整合，推出大数据相关产品和服务；同时也出现了 Splunk、Clustrix、Junar、DataSift 等大数据新兴公司，大数据市场正在积蓄巨大的商机。据 IDC 报告预测，大数据技术与服务市场将从 2010 年的 32 亿美元攀升至 2015 年的 169 亿美元，年增长率高达 40%，是整个 IT 与通信业增长率的 7 倍，将成为新的经济增长动力。

引领全球，开启大数据时代

美国推出“大数据研究和发展”计划之后，日本政府重新启动曾在日本大地震

后一度搁置的 ICT 战略研究，将重点关注大数据应用，并且联合国也随后发布了《大数据促发展：挑战与机遇》白皮书，全球范围内对大数据的关注达到了前所未有的热度，各类计划如雨后春笋般纷纷破土而出，大数据革命风起云涌。

2012 年 5 月，日本总务省信息通信政策审议会下设的 ICT 基本战略委员会召开会议，战略委员会大数据研究主任、东京大学教授森川博之强调，美国在大数据技术上处于领先地位，其 Google、Amazon 等网络企业在大数据应用领域有很强的优势，日本非常有必要在大数据方面制定综合性的战略。随后，日本文部科学省在 7 月发布了以学术云为主题的讨论会报告，指出为迎接大数据时代学术界面临的挑战，将重点推进大数据收集、存储、分析、可视化、建模、信息综合的各阶段研究，构建大数据利用的模型。

2012 年 7 月，联合国在发布的《大数据促发展：挑战与机遇》的白皮书中指出大数据时代已经到来，大数据对于联合国和各国政府都是一次历史性的机遇，报告讨论了如何利用大量丰富的数据资源帮助政府更好地响应社会需求，指导经济运行，并建议联合国成员国建设“脉搏实验室”，挖掘大数据的潜在价值。印度尼西亚在其首都雅加达建立的“脉搏实验室”由澳大利亚提供资助，于 2012 年 9 月投入运行。此外，乌干达也率先在其首都坎贝拉建立了“脉搏实验室”。

目前，大数据仍属于一个新兴前沿的概念，中国尚未从国家层面明确提出大数据的相关战略，但 2011 年 11 月中国工业和信息化部发布的物联网“十二五”发展规划中，提出四项关键技术创新工程，分别是信息感知技术、信息传输技术、信息处理技术和信息安全技术，其中的信息处理技术包括了海量数据存储、数据挖掘、图像视频智能分析，而这些技术都与大数据密切相关。同时，以广东为代表的地方政府率先启动大数据战略，坚持开放共享，推动大数据发展和社会创新。

大数据作为国家核心资产，是国家之间新的竞争焦点。在大数据领域的落后，意味着失守产业战略的制高点，意味着数字主权无险可守，意味着国家安全将在数字空间出现漏洞。相信在美国政府“大数据研究和发展”计划的影响下，欧盟、中

国等大型经济体在不久的将来必将出台引导性、倾斜性的政策，以抢占大数据的战略制高点，围绕大数据的新一轮竞争呼之欲出。

事实上，历史上曾经上演过类似的一幕，1993 年美国 “信息高速公路”计划一出台就引起了各国的强烈反应。日本政府在 1993 年 6 月发布拟建大规模超高速“研究信息流通新干线”计划，决心通过高速通信线路将全国研究机构与大学连接起来，并于 1994 年 5 月前后提出了日本版“信息高速公路”计划——《通信基础结构计划》和《通向 21 世纪智能化创新社会的改革》两个报告，决定分成三个阶段逐步实施网络建设。欧洲也不甘落后，在 1993 年 6 月哥本哈根欧盟首脑会议上，欧盟主席德洛尔首次提出“构建欧洲信息社会”的倡议，随后在同年 12 月，欧盟发布了旨在“振兴经济，提高竞争能力和创造就业机会”的白皮书，白皮书中明确提出构建欧洲版“信息高速公路”的设想，并成立了专门的工作小组负责计划的推进。此外，加拿大、韩国、新加坡等发达国家为争夺高新技术的发展优势，迎接 21 世纪的发展挑战，也都纷纷选择立即跟进，投入巨额资金，推出各自国家的“信息高速公路”计划，在全球范围内掀起一场热潮。

大数据本质上是人类社会数据积累从量变到质变的必然产物，是在“信息高速公路”基础上的进一步升级和深化，对人类社会的发展具有极其重大的影响和意义。根据现有形式，笔者完全可以断言，全球性的大数据发展浪潮即将到来。

第四节　Splunk 上市的影响

Splunk 在美国成功上市

美国软件公司 Splunk 于 2012 年 4 月 19 日在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低靡、股市持续震荡的大背景，Splunk 上市首日的突出交易表现尤其令人印象深刻。Splunk 在最初提交 IPO 申请确定的发行价区间为 8～10 美元，发行 1350 万股股票，计划募集资金 1.22 亿美元，以

发行区间中间值计算，公司市值为 9.94 亿美元。此后，Splunk IPO 的发行价一再被提高，从原来的 8～10 美元提高到 11～13 美元，后又提高到最终的 17 美元，融资规模不断扩大。然而好运并未止于此，Splunk 上市首日的开盘价即为 32 美元，开盘后震荡走高，以 35.48 美元收盘，市值超过 30 亿美元，首日即暴涨了一倍多，较最初提交 IPO 申请确定的发行价更是增长两倍多。Splunk 上市首日优异的交易表现超过了同年 3 月上市的移动广告网络公司 Millennial Media Inc.上市首日 92%的涨幅，与自 2004 年谷歌公司上市以来规模最大的互联网上市公司 LinkedIn Inc.的上市首日涨幅相当。

虽然曾经做过盈利承诺，但由于 Splunk 将大量的资金用于加快业务的增长，因而尚未实现盈利。根据 Splunk 公布的运营业绩显示（公司以每年 1 月 31 日前 12 个月作为一个报表年度），该公司在截止到 2012 年 1 月 31 日这一财年中亏损了 1099 万美元，较上一财年的 380 万美元有所扩大；但其营业收入正在快速增长，较上一财年增长了 83%，达到 1.21 亿美元，且该公司在最近几年的毛利率一直保持在 90%左右。

Splunk 虽然没有盈利，但在上市首日仍然受到资本市场的追捧，创造了卓越的成绩，这是自 2000 年互联网科技泡沫破裂以来，难得一见的盛事。这样的表现与 Splunk 迅猛的增长业绩有一定关系，但更为重要的是受到技术热点“大数据”的影响。大数据正成为继电子商务、物联网、云计算、移动互联等概念之后，被社会各界广泛追逐的新概念。Splunk 的成功上市，是产业界发展的一个重要里程碑，正式标志美国 IT 业开启了大数据元年。

Splunk 商业模式

Splunk 是一家领先的提供大数据监测和分析服务的软件提供商，成立于 2003 年，总部位于美国旧金山，在全球设有 8 个办事处，拥有 500 多名员工。Splunk 软件是一种高扩充性且通用的数据引擎，通过收集和索引由网络、应用程序以及移

动设备等不同来源和格式的机器数据，使用创新的数据架构来联机建立动态的创新的主题，允许用户不需理解数据的结构便可监控、检索、分析、图示化实时及历史机器数据流，帮助个人和组织实时分析数据，在各个方面提高运营效率，获得洞察力，并最终做出准确的判断和决策。

Splunk 的业务迎合了大数据时代企业对数据应用的需求。面对日益爆炸式增长的数据，企业需要对大数据进行处理，挖掘其中的潜在价值，以便能够有效地进行信息应用管理、IT 运营管理，增强整个公司与组织的洞察力。Splunk 的客户主要是财富 100 强公司，目前有来自 75 个国家 3700 多个客户在使用 Splunk 的产品和服务，客户所在的行业覆盖了教育行业（如哈佛大学、纽约大学）、金融服务行业（如美国银行、JP 摩根）、零售行业（如 Freshdirect、梅西百货）和高科技行业（如思科、摩托罗拉）等。

在大数据时代，企业对海量复杂数据的快速收集、存储、分析和管理的需求迅速膨胀，Splunk 作为大数据处理分析的代表企业具有十分强大的生命力和广阔的市场发展前景。

Splunk 上市点燃资本界和产业界的大数据热情之火

Splunk 的成功上市引发了风投对大数据行业的关注。大数据在零售、制造、医疗等领域都具有极大的应用价值，市场前景广阔。对于投资界而言，Splunk 的成功上市树立了大数据行业的榜样，风投开始加大对大数据行业的关注和投资力度，风险投资公司 Accel Partners 甚至发起了一个大数据公司投资基金。

数据分析领域的新星 DataSift 在 2012 年 5 月融资 720 万美元，投资方为 GRP Partners 和 IA Ventures。DataSift 的主要业务是提供 Twiitter、YouTube 及 Facebook 等社交网络历史和实时的信息结构化分析。

大数据处理创业公司 10gen 在 2012 年 5 月融资 4200 万美元，投资方包括 New Enterprise Associates、Sequoia Capital、 Flybridge Capital Partners

和 Union Square Ventures。10gen 的主要产品是非关系型数据库 MongoDB，付费用户已达到 500 家，行业覆盖金融、通信和媒体等。

企业云存储服务公司 Nirvanix 在 2012 年 5 月再融资 2500 万美元，由风投公司 Khosla Ventures 牵头，Valhalla Partners、Intel Capital、Mission Ventures 及 Windward Ventures 等投资公司共同参与。Nirvanix 为客户提供数据的存储、传输以及处理服务，并在巨量、非结构化的数据存储方面处于领先地位。

将大数据应用于医疗保健行业的初创企业 Predilytics 在 2012 年 9 月 A 轮融资中获得了 600 万美元，资金提供方为 Flybridge Capital Partners、Highland Capital Partners 和 Google Ventures。Predilytics 运用大数据、机器学习技术持续地分析结构型和非结构型的客户数据，为医保领域提供洞察力。

大型数据处理公司 Attivio 在 2012 年 10 月获得了 3400 万美元的首轮融资，由 Oak Investment Partners 领投。Attivio 公司的核心产品是一个智能引擎——AIE（Active Intelligence Engine），通过整合结构化和非结构化的各类数据，为客户提供智能的搜索和分析服务，目前 UBS、Cisco 和德国电信等都是 Attivio 的客户。

Splunk 上市后主要大数据公司融资情况见表 2-1。

表 2-1 Splunk 上市后主要大数据公司融资情况

公司名称	融资时间	融资额/万美元	轮次
DataSift	2012 年 5 月	720	B
Evernote	2012 年 5 月	7000	D
Junar	2012 年 5 月	120	A
Precog	2012 年 5 月	200	C
10gen	2012 年 5 月	4200	E
Nirvanix	2012 年 5 月	2500	C
Precog	2012 年 5 月	200	A
Mixpanel	2012 年 5 月	1025	A
Sumall	2012 年 6 月	150	A
Delphix	2012 年 6 月	2500	C

续表

公司名称	融资时间	融资额/万美元	轮次
Clustrix	2012 年 7 月	675	B
GoodData	2012 年 7 月	2500	B
ParStream	2012 年 8 月	560	A
InsideSales	2012 年 8 月	400	A
TalentBin	2012 年 9 月	1000	A
Predilytics	2012 年 9 月	600	A
Datameer	2012 年 9 月	600	C
Trifacta	2012 年 10 月	430	A
Ngdata	2012 年 10 月	250	A
RainStor	2012 年 10 月	1200	C
DataStax	2012 年 10 月	2500	C
Attivio	2012 年 10 月	3400	A

资料来源：互联网资料、国金证券研究所。

围绕大数据的投资囊括了大数据的收集、存储、搜索、分析处理、可视化以及行业应用等整个产业链。数据背后存在着巨大的市场和价值，那么（在利益的驱使下）大数据技术将日益成熟。

Splunk 上市促使 IT 厂商加快大数据布局

面对前景广阔的大数据市场，IT 厂商纷纷排兵布阵，或发布战略，或推出产品，或实施并购，或开展合作，总之动作频频，意图在这巨大的蛋糕中分一杯羹。

EMC 实施三步曲战略，构造数据星球

EMC 在 1979 年成立于美国，是一家全球领先的信息存储及管理产品、服务和解决方案的 IT 厂商，2011 年的营业收入为 200 亿美元，净利润为 25 亿美元。EMC 在 2012 年 5 月底召开的 EMC World 大会上，一口气发布了 42 款新技术和新产品，在 2011 年提出的“大数据恰逢云计算”的基础上更精进一步，提出了“数据星球”的概念。至于如何构建数据星球，EMC 的云和大数据战略无疑将提供支撑。

在大数据时代，EMC 将自己的使命确定为引导客户和合作伙伴的大数据之旅，帮助他们利用大数据机遇加速业务转型。未来三年的大数据业务发展目标设定为每年翻一番，同时 EMC 将大数据发展战略分为三个阶段：第一阶段是构建云基础架构，利用 EMC Isilon 和 EMC Atmos 两个产品解决快速增长的复杂的海量数据的存储问题；第二阶段是提供数据科学协作和自助服务，也称为社交化阶段，EMC 的 Greenplum Chorus 便是一个社交化的数据处理平台，既可以处理结构型和非结构型数据，又可以帮助数据库管理员、数据库分析师、工程师等人员实现团队协作与分工；第三阶段是提供实时决策支撑，实现数据“货币化”，EMC 在 2012 年 3 月收购的 Pivotal Labs 能够帮助客户快速构建大数据应用。

IBM 大数据战略全面升级

IBM 在 1911 年创立于美国，总部位于纽约州阿蒙克市，2011 年营业收入为 1069 亿美元，净利润近 159 亿美元，是全球最大的信息技术和业务解决方案公司。

在大数据时代，IBM 积极应战。从 2012 年年初提到大数据，到 5 月发布智慧分析洞察“3A5 步”动态路线图，再到 9 月举办“大数据 · 大洞察 · 大未来”发布会，IBM 通过内部资源的全面整合，搭建了集软件、硬件和服务为一体的大数据平台，宣告大数据战略全面升级。其主要体现在三个方面：一是“全面的战略理论”——“3A5 步”，即掌控信息（Align）、获悉洞察（Anticipate）、采取行动（Act）、学习和转型；二是“全面的解决方案”，主要包括 Hadoop 系统、流计算（Stream Computing）、数据仓库（Data Warehouse）和信息整合与治理（Information Integration and Governance），其中信息整合与治理是 IBM 独有的技术，代表产品为 Optim 和 Guardium；三是“全面的落地实践”，IBM 在制造、电信以及金融等行业积累了丰富的经验。

阿里巴巴高度重视数据业务

阿里巴巴集团在 1999 年成立于中国，是中国最大的电子商务公司，目前旗下

拥有阿里巴巴 B2B、淘宝网、天猫商城、聚划算、一淘、中国雅虎、阿里云、中国万网、一达通、CNZZ 等众多公司。阿里巴巴集团经过十多年的发展，平台上积累了大量的数据，且这一数据仍在快速增长。

为挖掘大数据的价值，2012 年 7 月阿里巴巴集团在管理层设立了“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出了大型的数据分享平台——“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。马云强调：“假如我们有一个数据预报台，就像为企业装上了一个 GPS 和雷达，你们出海将会更有把握。”因此，阿里巴巴集团希望通过分享和挖掘海量数据，为国家和中小企业提供价值的信息。

为迎接大数据时代的变革与机遇，除了 EMC、IBM 和阿里巴巴，甲骨文、微软、SAP、惠普、英特尔、百度等传统 IT 巨头也在积极布局大数据，挖掘大数据中蕴含的“金矿”，一时间大数据市场热闹非凡。

第五节　数据科学与信息产业大会的召开

2012 年，大数据主题风靡各种论坛、会议，无论是学术界还是产业界，大家纷纷探讨大数据带来的影响和变化。在众多的会议当中，笔者认为“数据科学与信息产业大会”具备里程碑的意义，因此用一节的篇幅来特别说明。

2012 年 11 月 17 日，由中国科学院院士鄂维南老师召集，来自国内外学术界的泰斗和前线科研人员、产业界高管共聚一堂，召开了“数据科学与信息产业大会”，明确地提出了“数据科学”这一新兴学科。与会的中科院院士有张恭庆、石钟慈、李大潜、林群、马志明、郭雷、袁亚湘，工程院院士有崔俊之、李国杰、高文。

这次会议的与众不同之处是，学术界敞开大门广泛邀请在数据领域有产业需求的公司参会。参加此次大会的既有互联网经济的代表，也有新兴科技公司和老牌的软件企业，包括京东商城、百度、用友、拓尔思、启明星辰等多家公司，反映出学

术界一股清新、务实之风。鄂院士在发言中谈到，“科学研究最重要的一环是提出前瞻性的问题。提不出问题，就只能跟在别人后面，走一条从文献到文献的路子。对我国的科技界来讲，在很多学科，由于来自实际应用领域的限制，提出前瞻性问题的确是件很困难的事情。但数据科学则不然。由于我国人口众多这一特殊情况，和我们特殊的文化、文字、历史背景和社会发展的需要，我们在数据科学领域的很多问题自然就是前瞻性的。关键是我们能否用前瞻性的方法去面对这些问题。如果做好了这一点，我们在数据科学领域就自然而然地走到了世界的前沿。”

会议同时邀请资本市场人士参与。资本市场以其特殊的敏感性，从大数据刚刚发端，就对此主题保持高度的关注。“数据科学与信息产业大会”汇集了学术界、产业界和资本市场的顶级资源。大家都期待在促进学术进展、产业繁荣方面，做出更多的贡献。

第六节　大数据创新的策源地——云基地大数据实验室

北京云基地以“基金+基地”的模式建立中国云计算的生态系统，构建全球领先的立足于中国云计算产业的企业群落。大数据实验室通过投资与研究两轮驱动，同时带动市场创新与科研创新，通过灵活的机制吸引与聚集了大量人才、数据与市场机会，提出新的科学问题，孵化新的创业公司。

2012 年开始兴起的大数据热潮得到了上到国家领导人下到中小企业主、普通老百姓的广泛关注。全社会都意识到，大数据存在巨大价值，大数据的出现将对人们的生产、生活带来巨大的改变。大数据涉及新的技术、新的业务模式、新的决策流程、新的思维模式……总而言之，释放大数据价值的关键在于创新。

历史上的创新从何而来？很多人以为创新是某个离群索居的发明家的奇思妙想，这些发明家只是出于对知识的追求而从事发明工作，而不考虑其发明的经济回报。如果纵观对人类产生重大影响的创新出现的历史可知，其不尽然。随着社会环

境的变化，主导不同时代的主流创新模式也会有所不同。例如，达芬奇、哥白尼、伽利略等天才的单打独斗式的创新，在人类创新史的早期由于信息交流不发达、市场力量不成熟占据创新的主流地位。在这个时期，这些天才们独立发明了凹透镜、地球仪、日心说。随着 17 世纪古登堡印刷术在西方的普及，科学的思想能够更廉价地进行存储与传播，同时教育系统也随之发达，这使得天才们能够相互激荡，形成思想的网络。围绕着剑桥大学、英国皇家学会等科学团体与科学网络，人们对自然的探索达到了一个前所未有的高度，万有引力定律的发现、光合作用的发现、林奈分类法的发明、望远镜的发明，都是这个时代的代表性产物。同时，随着科学的发展与市场的成熟，能够带来巨大商业价值的蒸汽机、多轴纺纱机、摆钟等发明也在创新的网络中被培育出来，这些发明引发了工业革命的爆发。自 19 世纪以来，大学与市场成为驱动创新的两股相互交织的重要力量。在大学与科研机构，天才们发明了元素周期表、阿司匹林、青霉素、全球定位系统（GPS）等，发现了电子、细胞分裂等。在市场，天才们发明了飞机、电话、汽车、洗衣机、计算机等。这些发明都不是天才们单打独斗的结果，在创新的竞技场上，天才们交流思想、相互合作、相互竞争，运用集体的智慧创造了人类的奇迹。虽然，在科学研究领域，某些天才的独立思考还能够对整个世界产生重要的影响，他们发明了相对论，发现了双螺旋结构、X 射线。但是，在市场，远离创新网络的个人努力取得重大成就的案例却鲜有发生。

中国在以 17 世纪、19 世纪的创新竞赛中落后于西方，这使得中国的社会与经济长期处于被动地位。中国向来不缺少聪明且勤奋的天才，缺乏的是催生创新的土壤。21 世纪开始的大数据时代，中国第一次和西方站在同一条起跑线上。中国有可能根据中国特点，利用制度创新，围绕市场、数据与人才构建大数据创新网络与生态系统，使得中国的数据科学与商业在世界上占有一席之地。这个创新网络一方面将促进科学研究，使得研究者能够基于现实世界的海量数据提出与解决具有国际领先水平的科学问题。另一方面，这个创新网络将实现人才、数据与市场的协同效应，

孵化与催生具有国际竞争力的大数据企业。另外，由于人才、数据、市场的相互促进，国计民生的改善将对这个创新网络有越来越强的依赖，进而使其成为整个社会的大数据基础设施。

虽然这样的愿景看起来很美，但是构建这样的创新网络需要灵活的商业模式与合作机制，需要人才的聚集、数据的聚集、市场的聚集等诸多条件的相互作用，诸多条件的汇聚颇为不易。依托于北京云基地的大数据实验室（下文简称“大数据实验室”）因其独特的优势而构建大数据创新网络。大数据实验室的大数据创新将打破市场创新与高校创新的界限，形成新的创新范式。大数据实验室以投资启动与加速大数据创新进程；大数据实验室依托北京云基地的业务合作能力为大数据创新提供市场与现实世界问题；大数据实验室以其资源整合能力及数据处理能力为大数据创新提供数据基础；大数据实验室以优秀的人才资源吸引更多的大数据人才合作；大数据实验室依托北京云基地的云计算能力为大数据创新提供所需的技术支持与计算能力。它必将成为中国大数据创新的策源地。

市场

大数据是利用数据或信息为决策服务的，是用来提升企业或者个人的决策效率的。大数据利用了科学的方法论，但它绝不是一门象牙塔里的学问。如鄂维南院士所言“科学研究最重要的一环是提出前瞻性的问题。提不出问题，就只能跟在别人后面，走一条从文献到文献的路子。”大数据创新更需要提出有现实意义的前瞻性问题。有意义的大数据创新问题的提出必然来源于现实社会的企业与个人，来源于现实世界的人们利用数据解决现实世界问题的需求。然而，传统的科研机构的设置使得研究者与对现实世界的问题具有切身感知的运营者之间存在一定距离。为了满足大数据创新的需求，需要研究者能够和现实世界的业务需求相互靠近。

另一方面，对创业者而言，大数据创业也与传统的 IT 产品有所不同，大数据创

业的过程也是一个数据探索与发现的过程。大数据创业虽然也能事先规定产品的方向、所要解决的业务问题的业务领域，但是数据助力业务领域决策的方式及限度将随着创业者对数据认识的深入而变化。例如，待解决业务问题本身的不确定性或可预测性存在一定限度，而其限度究竟在哪里，只有通过对数据的探索才能获知。随着对数据的探索与理解的深入以及随着对业务问题的理解与深入，大数据创业者有可能提出创新的业务问题或提出新的方法解决既有的业务问题。所以，大数据创业的过程，本身也是对数据能力与界限探索的过程。大数据创业更需要以快速试错、快速迭代为特点的精益创业（Lean Startup）方法论的指导。而这要求大数据创业者能足够贴近最终用户，使得其大数据产品与用户的需求以及决策流程能快速配合。这也需要大数据创业者与业务需求的足够靠近。

依托于北京云基地的大数据实验室将是拉近大数据研究者、大数据创业者与现实世界的业务问题与业务应用进而推动大数据创业的促进者。通过大数据实验室，大数据领域的创业者与研究者将有机会和各行业有代表性的企业直接沟通，理解真实世界的现实问题，并有机会将其创新直接在这个合作网络的企业或企业的客户中进行测试与验证。同时，大数据实验室与某些行业的领导者进行合作，专门组织力量建立社区，以孵化与研究用于提升该行业产业价值的大数据与技术。

数据

影响大数据创新的另一个关键要素是数据。没有数据的大数据创新是无源之水，无本之木。然而，大数据创新过程中的数据访问与数据处理存在一定的门槛，这妨碍了大数据的创新。其主要表现在以下三个方面：

其一，一些大型企业掌握了海量的高价值数据，而这些数据中往往包含敏感信息或隐私信息，在现阶段数据立法尚不完善的情况下，企业为了保护数据安全与数据隐私，往往禁止第三方对数据的访问。然而，数据中还往往包含能够促进国计民

生的高价值信息，完全禁止第三方的数据访问有因噎废食之嫌。应该在可控、授信、全程监督的情况下，鼓励对大数据价值的探索，这有助于大数据价值的实现。大数据应用的建立一般分为建模（数据探索）和应用两个阶段，建模（数据探索）阶段可能会访问某些敏感信息，但是在应用阶段则未必需要将这些信息对外发布。大数据实验室凭借其在中国国内的地位与信用，和一些大型企业展开合作。在大型企业的授权与监督下，大数据实验室的研究者有机会对这些数据进行探索，而这是包括外企在内的很多大数据研究者与应用者所不能获得的机会。

其二，如果按照传统商业智能的观点，数据挖掘的工作中 80%左右的工作量用于数据准备工作。与大型企业所掌握的高价值数据不同，目前大数据创业者与研究者有机会通过互联网等渠道免费获得一些有价值数据，或者以付费的形式购买一些公开的有价值数据。然而，数据的搜集、整理需要耗费大量的工作量。由于这些数据能够通过公开渠道获得，数据的搜集与整理往往只是构成这些创业者与研究者的成本，而不能为其带来核心价值。大数据实验室通过自主开发、合作等形式获得与维护大部分大数据创业者与研究者所需的公用数据集合，以降低大数据创业者与研究者用于数据搜集与维护所需的成本，让大数据创业者与研究者集中精力于更能带来价值的数据挖掘与应用开发工作中。

其三，以海量的非结构化数据为代表的大量数据的产生是大数据时代的一个重要特点。对海量的、非结构化的数据进行处理，一方面需要大量的计算资源，另一方面需要专门的数据处理技能。这从资本投资以及专业技能投资两个方面对大数据创新与创业设置了门槛。然而，北京云基地进驻的云系企业，都是分布在云计算产业链各个环节中的行业主导企业，聚集了一大批国内外云计算人才。其产品和服务涵盖云计算多个环节，包括服务器、数据集装箱、瘦终端等硬件产品的设计和生产，云中间件、云管理平台、桌面虚拟化等基础软件研发；智能知识库、分布式计算、

视频云等应用软件，以及定制化的云计算解决方案，构成完整的云计算产业链。依托于北京云基地的大数据实验室能够为大数据创新与创业提供大数据处理所需的软硬件平台与专业人才。同时，北京云基地与北京市中关村管委会共同推动大数据产业联盟的建立与发展，使得大数据创新者能够有机会接触到更多的大数据创新所需资源与专业信息技术。

人才

大数据时代最稀缺的资源是人才。谷歌的首席经济学家 Hal Varian 曾经说过，未来最性感（sexiest）的职业是统计学家。大数据人才的招募、培养与使用将是大数据创新与创业所面临的最大挑战。通过合理的模式释放大数据人才价值的过程同时也是释放大数据价值的过程。大数据创新人才缺乏的原因多种多样，而对其破解的方式也多种多样，大数据实验室在此做了有益的尝试。

很多企业拥有很好的数据基础以及很有意义待解决的数学问题，且拥有引入大数据战略的强烈愿望。但是由于企业规模或企业发展现状所限，企业没有机会接触到最好的数据科学家，只有将其大数据战略暂时搁置。大数据实验室坐落于中国智力资源最为密集的北京中关村地区，同时在上海高校密集的杨浦区设有分部，这使得大数据实验室能够与中国顶尖的数据科学家进行合作。通过与大数据实验室的合作，有大数据方面抱负的企业或者创业者有机会求助于中国顶尖的数据科学家以解决关键的大数据难题。而大数据实验室所维系的多样化的数据科学家网络能够与大数据实验室共享，这也解决了对技能多样化的需求以及人才成本分摊的问题。而对于数据科学家而言，因为问题具有多样化的特点，且具有现实意义，这也对这些数据科学家产生强烈的吸引力。

除了数据科学家之外，大数据的创新与创业需要多种具有专门知识与技能的人

才加盟或辅助。这些技能包括 IT 能力、行业知识、创业知识、投资知识等。大数据实验室建立并维系了具备多种专业技能的导师（mentor）网络。这些来自各行业的专家定期或按需对大数据创新者进行指导，以提升创新者的创新成功率。

另外，大数据实验室也为大数据人才提供了多种可能性，使其能够根据自身特点与偏好选择合适的模式施展其大数据才华。对大数据研究感兴趣的大数据人才，可以选择和大数据实验室开展合作研究，研究数据、探索大数据相关技术。同时兼具商业才能与技术才能的大数据人才，可以选择获得大数据实验室的资助开展创业活动。

大数据实验室以灵活的模式与大数据人才展开合作，使大数据人才能够各展所长。人才价值的实现是数据价值实现的前提与必要条件。

导读：

1. 数据资产是产业兴衰的关键因素。新兴的公司凭借独一无二的数据资产，不断扩张商业版图，持续侵袭传统产业的领地。没有人能遏止它们的扩张势头，除非其也有庞大的数据资产，并能善加运用。谷歌、亚马逊、Facebook 是新兴公司的典型代表。
2. 正确评估数据资产的潜在价值，是公司估值和投资判断的重要组成部分。数据资产评估模型，从规模、活性、维度、颗粒度、关联度五个方面给出判断原则和指引。
3. 数据资产的巨大价值，通过不同的商业模式发挥到极致。大家在文章中可以看到，因为有了数据资产，原来无法开展的商业模式变得简单、可靠，原来无法服务的用户变得触手可及，原来难以捉摸的未来趋势变得清晰可循。

第三章

数据成为资产

数据资产是产业兴衰、企业存亡的关键因素。

——笔者

就像厚厚的沉积岩忠实地记录了不同世代的沧海桑田一样，大数据封存了人类社会的共同记忆。普罗大众们无缘在史书中占据哪怕是立锥之地，但是他们每一个人都在大数据中鲜活永久的存在。现在的人们只能凭着《清明上河图》等为数不多的绘画珍品和一些史书来推断历史上的社会百态，而我们的后代却可以通过大量的照片、视频、个人博客等素材来再现社会任意的横断面。从这个角度来看，大数据社会意义之深远，甚至超过对当下产业的启迪。这方面的研究和著述交给社会学家系统分析，笔者依然聚焦在大数据为产业带来的价值。

长期以来，经济学著作中，土地、资本和人力并称为企业的生产要素。人类进入工业时代以来，技术成为独立的生产要素之一。“上九天揽月，下五洋捉鳖”，离开技术的发展，是难以想象的。但是在信息时代，数据将成为独立的生产要素。有人把“数据”比喻为工业时代的石油，事实上 “数据”和农耕时代“土地”的属性更加接近。如果企业拥有某类相对完整、全面的数据，退可偏安一隅，进可跃马中原。

互联网领域，令人称道的谷歌、亚马逊和 Facebook，分别拥有不同的数据资产。谷歌之所以能打破微软垄断的铁幕，依仗的就是世界上最大的网页数据库，并建立了充分发挥这些数据资产潜在价值的数字媒体商业模式。许多公司开始把谷歌当作竞争对手，依葫芦画瓢推出和谷歌类似的搜索引擎，但是，包括微软公司在内没有一家可以撼动谷歌的根基，直到 Facebook 推出 graph search 引擎，才让谷歌感到真正的威胁。原因很简单，Facebook 拥有谷歌缺乏的一类数据资产——人们的关系数据，这是 Facebook 区别于所有竞争对手的关键因素。当谷歌和 Facebook 打得不可开交的时候，亚马逊却乐得坐山观虎斗。因为无论是谷歌还是 Facebook 都可以帮助亚马逊卖出更多的商品。亚马逊拥有世界上最大的商品电子目录。当所有公司对苹果的平板电脑横扫世界束手无策的时候，亚马逊庞大的商品帮了大忙，人们愿意购买亚马逊的平板电脑，因为可以免费获得海量的图书。和亚马逊相比，缺少电子图书，恰巧是苹果的弱项。所以没有独一无二的数据资产，几

乎无法参与巨人间的游戏。

中国的互联网市场也是硝烟弥漫。阿里巴巴旗下的一淘网，抓取京东商城的客户评论数据；京东则采取技术手段屏蔽一淘的爬虫。另一方面，电商则纷纷抓取竞争对手的各类商品的实时价格，作为评估对手战略动向、促销战术的重要依据。这还只是在表面现象，事实上互联网平台型的公司，都在围绕数据资产为核心整合产业生态。它们推出新的产品、新的服务，就会收集更多类型的数据。数据越多，不同类型数据之间的关联性、实时性越强，就会提炼出更有价值的信息，指导它们开展各类精准的广告业务、金融业务。马云在 2012 年网商大会上，鲜明地提出阿里巴巴未来的战略是围绕三大方向即平台、金融、数据展开。平台汇聚数据，数据衍生金融，金融反哺平台。可见互联网公司对于数据资产的战略价值，认知最为深刻，行动最为果断。京东商城也已经启动供应链金融服务。表面上看，电子商务公司和金融机构井水不犯河水，其实电商凭借数据积累，已经侵入到金融行业的腹地。

在电信行业，数据已然成为推动运营商整体转型的战略性资产。常常听到各种唱衰运营商的声音，OTT 业务就是许多人不看好运营商的重要理由。OTT(Over the Top)，意为“过顶传球”，通信业用这个词来形象地比喻微信、Skype 等互联网应用。这些应用利用互联网传递语音，降低了传统通话业务的使用时间，进而延伸到各类使用运营商基础网络通信功能，但是却会消弱运营商话语权的业务。业内著名咨询顾问公司 Ovum 预测，到 2020 年，OTT 类语音服务将让全球电信业累计损失 4790 亿美元，占语音业务收入总额的 6.9%。更有人悲观的预测，全球性的基础语音通话免费将是大势所趋，届时运营商何去何从?

事实上，电信运营商掌握的数据，令人垂涎三尺。第一，这些数据都是实名产生的，可以具体到每一个消费者；第二，通过这些数据可以直接获取人们精确的位置信息；第三，仅仅利用这些数据，可以精确地获悉人们的生活起居、行为爱好等。谷歌公司的业务模式，就是建立在对这些数据的分析挖掘基础之上的。因而，运营商实际上坐拥“金山”只是“敢不敢”或者“能不能”开采的问题。第一个问题主

要是法律的限制，如何善用这些数据“不作恶”，也就是不能侵犯用户的隐私。目前数据资产归属权和使用权界定不清晰，是导致运营商在开挖金山时畏手畏脚的主要原因之一。第二个则是人才和机制方面的问题。这个说起来话长，留给运营商们自己去探讨吧。但是，电信业门口的“野蛮人”，恐怕不会留给运营商多少从长计议的时间。谷歌公司已经在提供基础的电信业务了，其用意就是要在信令级获取人们的行为数据。

在金融行业，对数据资产的争夺，已经关乎金融产业的未来格局。金融行业自其诞生以来，就是靠信息驱动的行业，所以金融业内的公司从不吝惜在信息技术方面的投资。但是互联网发展之迅速，还是令金融业有些措手不及。Facebook、腾讯等大型的互联网帝国都在发行虚拟的“货币”。这些“小打小闹”似乎还没有触及银行传统借贷业务的核心，但是电商突然杀入小额贷款、供应链金融等领域，却让银行感受到了什么是切肤之痛。更令银行难堪的是，离开电商提供的中小企业交易数据，银行缺少可靠的数据来源去分析众多客户的经营风险。得中小企业者，得天下；得数据者，得中小企业。银行已经在这场事关未来的“数据资产”争夺战中，落了后手。各大行业数据资产价值对比见表 3-1。

表 3-1　各大行业数据资产价值对比

	银行	运营商	社交网络	电子商务	搜索引擎
数据标识	账号	手机号码	邮件账号	注册账号	无
身份真实性	●	●	◒	●	◔
规模	●	●	●	●	●
颗粒度	●	●	●	●	●
活性	◔	●	●	●	●
多维	◔	◔	●	●	●
关联性	◔	●	●	◔	◔

●高　◔低

政府拥有的数据，则更加全面、翔实，反映一个国家、社会的各方面。譬如，所有人和法人的账户都会在人民银行备案，每个人的身份数据则在公安部的电脑中

存档，海关则会忠实地记录每天进出口的货物、人流……这个可以一直写下去，但这些数据资产处于“数据割据”的状态，难以发挥聚合的效应。数据割据状态，纵向上体现为上级单位无法全面实时访问下级单位的详细数据；横向上体现为部门间的利益纠葛，而不希望、不愿意把数据开放给其他部门。随着技术的进步，各部委逐步实施“数据集中”项目，消除数据纵向割据现象。但是数据横向割据，则要靠法律或者行政的手段来克服。因为数据一旦整合，其发挥的价值将难以估量。

譬如，“综合治税”的问题。“涵养税源”是税务部门工作的最高境界，但是总有一些企业逃税和避税。据说奥巴马政府也在为苹果等大户避税的问题困扰，可见这是一个世界性的问题，并非“中国特色”。综合治税的重要抓手就是不同部委之间数据的交叉印证。企业用电、用水、报关，电商平台上的销售、采购等数据综合起来，理论上应该可以判断一家企业的营业收入。

第一节　数据资产价值及评估

提要：

1. 大数据思维的重要性远远超过数据资产，具备大数据思维，才能够积累数据资产；不具备大数据思维，则可能弃珍宝如敝屣。
2. 数据资产评估模型包括颗粒度、维度、活性、规模、关联度五个方面。
3. 忠实记录人们行为的一类数据，就像巫师的水晶球一样，具备洞悉未来的能力。华为公司据此数据可以提前 1～2 个月预测公司未来的收入；美国上市的首家大数据公司 Splunk，就是提供此类数据分析工具的公司。

本节的写作源于实际的需求。自笔者在资本市场提出大数据的概念以来，不断接到各种邀约，帮助评价一些公司的投资价值，包括二级市场、一级市场和初创的企业。更有一些大公司也在邀请笔者讲解如何理解大数据，交流中都会涉及到如何认识数据资产的价值问题。笔者根据以往的经验和判断依据，得出大数据资产价值评估模型。这个评估模型并非出于学术目的，只是为大家评价数据资产提供了思考的框架。随着笔者接触的企业越来越多，数据资产评价模型也在不断地修正和完善。

需要着重申明的是，公司最重要的是建立大数据思维，而非仅仅盯住数据资产。以电信运营商为例，每个人每次打电话都会产生一条通话记录，这些记录用大数据的视角来看，都是宝贵的资产。如果运营商弃之如敝屣，或者留着压箱底，保存在缓慢的磁带上，不善加利用，那么运营商的数据再多，都不能被称为“大数据公司”。

数据思维

有两个故事可以说明建立数据思维的重要性。第一个故事关于台塑集团的创始人王永庆，第二个故事的主角是林彪[①]。

王永庆被全球化工行业奉为经营之神，很多企业家都把他的管理经验当作最实用的教科书。16 岁的王永庆借款 200 旧台币，开始创业，经营米店。但是居民一般都有自己常去的店铺，而那些店铺也想尽办法来吸引老客户，所以王永庆的新店冷冷清清。王永庆在挨家挨户拜访客户的时候，发现买米的大多是家庭主妇，于是提出送米上门的服务。他总是认真地帮客户清洗米缸，把陈米清理出来，再把新米倒入米缸，这样保证客户不会一直积攒陈米。王永庆边劳动，边和主妇聊天，留意米缸的大小、家里的人口、发工资的日期等信息。回到店里，王永庆就会细心地把这些数据记录在小本上，日复一日，从不间断。王永庆根据这些数据，推算客户大约在什么时间需要新购大米，总是在客户购买之前，上门把新米倒入客户的米缸。

① 林彪数据素养的案例选自陆迪的博客，http://blog.sina.com.cn/s/blog_53f9871e0101ewbh.html。

王永庆的销售额开始大幅增长，从开始一天不足 12 斗的销量，到后来可以每天卖出 100 多斗。10 年的卖米生涯，奠定了他一生事业的基础。

1948 年辽沈战役开始之后，在东北野战军前线指挥所里面，每天深夜都要进行例常的“每日军情汇报”：由值班参谋读出下属各个纵队、师、团用电台报告的当日战况和缴获情况。

那几乎是重复着千篇一律的枯燥无味的数据：每支部队歼敌多少、俘虏多少；缴获的火炮、车辆多少，枪支、物资多少……

司令员林彪的要求很细，俘虏要分清军官和士兵，缴获的枪支要统计出机枪、长枪、短枪，击毁和缴获尚能使用的汽车也要分出大小和类别。

经过一天紧张的战斗指挥工作，人们都非常疲劳。整个作战室里面估计只有定下这个规矩的司令员林彪本人，还有那个读电报的倒霉参谋在用心留意。

1948 年 10 月 14 日，东北野战军以迅雷不及掩耳之势，仅用了 30 小时就攻克了对手原以为可以长期坚守的锦州并全歼了守敌十余万之后，不顾疲劳，挥师北上与从沈阳出援的敌精锐廖耀湘基团二十余万在辽西相遇，一时间形成了混战。战局瞬息万变，谁胜谁负实难预料。

在大战紧急中，林彪无论有多忙，仍然坚持每晚必作的“功课”。一天深夜，值班参谋正在读着下面某师上报的其下属部队的战报。说他们下面的部队碰到了一个不大的遭遇战，歼敌部分，其余逃走。与其他之前所读的战报看上去并无明显异样，值班参谋就这样读着读着，林彪突然叫了一声“停！”他的眼里闪出了光芒，问：“刚才念的在胡家窝棚那个战斗的缴获，你们听到了吗？”

大家带着睡意的脸上出现了茫然，因为如此战斗每天都有几十起，不都是差不多一模一样的枯燥数字吗？林彪扫视一周，见无人回答，便接连问了三句：

“为什么那里缴获的短枪与长枪的比例比其他战斗略高？”

“为什么那里缴获和击毁的小车与大车的比例比其他战斗略高？”

“为什么在那里俘虏和击毙的军官与士兵的比例比其他战斗略高？”

人们还没有来得及思索，等不及的林彪司令员大步走向挂满军用地图的墙壁，指着地图上的那个点说："我猜想，不，我断定！敌人的指挥所就在这里！"

随后林彪口授命令，追击从胡家窝棚逃走的那部分敌人，并坚决把他们打掉。各部队要采取分割包围的办法，把失去指挥中枢后会变得混乱的几十万敌军切成小块，逐一歼灭。司令员的命令随着无线电波发向了参战的各部队……

而此时的廖耀湘，正庆幸自己刚刚从偶然的一场遭遇战中安全脱身并与自己的另外一支部队汇合。他来不及休息就急于命令各部队尽快调整部署，为下一阶段作准备。可是好景不长，紧追而来的解放军迅速把他的新指挥部团团围住，拼命攻击，漫山遍野的解放军战士中，不断有人喊着："矮胖子，白净脸；金丝眼镜湖南腔，不要放走廖耀湘！"

把对方指挥官的细节特征琢磨到如此细微，并变成如此威力巨大的顺口溜，穿着满身油渍伙夫服装的廖耀湘只好从俘虏群中站出来，无奈地说"我是廖耀湘"，沮丧地举手投降。

廖耀湘对自己精心隐蔽的精悍野战司令部那么快就被发现、打掉，觉得实在不可思议，认为那是一个偶然事件，输得不甘心。当他得知林彪是如何得出判断之后，这位出身黄埔军校并留学法国著名的圣西尔军校，参加过滇缅战役，在那里把日本鬼子揍得满地乱爬的新六军军长说："我服了，败在他手下，不丢人。"

取得这场重要战役胜利的其中一个关键因素，居然出于获胜方的统帅夜半时分对一份普通遭遇战之后的战报的数据分析，这来源于他"从红军带兵时起，身上有个小本子，上面记载着每次战斗的缴获、歼敌数量"的优良军事素养。

数据的积累、挖掘、分析、归纳、整理，是一支优秀团队所必须具备的基本素养，没有它，你永远是匹夫之勇。

数据资产评估模型

优秀的数据思维，必然反映在优质数据资产。人们难以定量评价一个人的数据

思维，所以只好退而求其次，关心在数据思维的影响下，数据资产的优劣。数据资产的价值从五个维度来评估，分别是规模、活性、多维度、关联性、颗粒度，如图 3-1 所示。

这五个维度，没有绝对的数值可以参考。林彪和王永庆的“小本本”都非常有效，如果按现在的度量衡，他们的数据量估计连 1MB 都达不到。所以，在模型的定义中，仅仅给出定性的描述，具体到每个行业，需要根据这个模型来灵活运用。

颗粒度指标反映数据的精细化程度。那些宏观的数据，价值含量较低。相反那些细化到个人、单品的数据，才会带来前所未有的洞察力，这也是和精细化管理的思想紧密相关的。早期管理者认为工业产品没有差别，同一个批次、型号的产品是一模一样的。但是现在人们需要管理到“单品”，也就是每一件产品。提高社会治理水平，也是逐渐细化“管理单元”的过程。秦始皇设定“郡县”，这是当时最小的国家机构，其最高长官就是传统戏剧中经常戏谑的“七品芝麻官”。但是现代的管理单元已经细化到 100m × 100m 的正方形，形象地称为“网格”，一个网格中，很可能只有一座楼房而已。

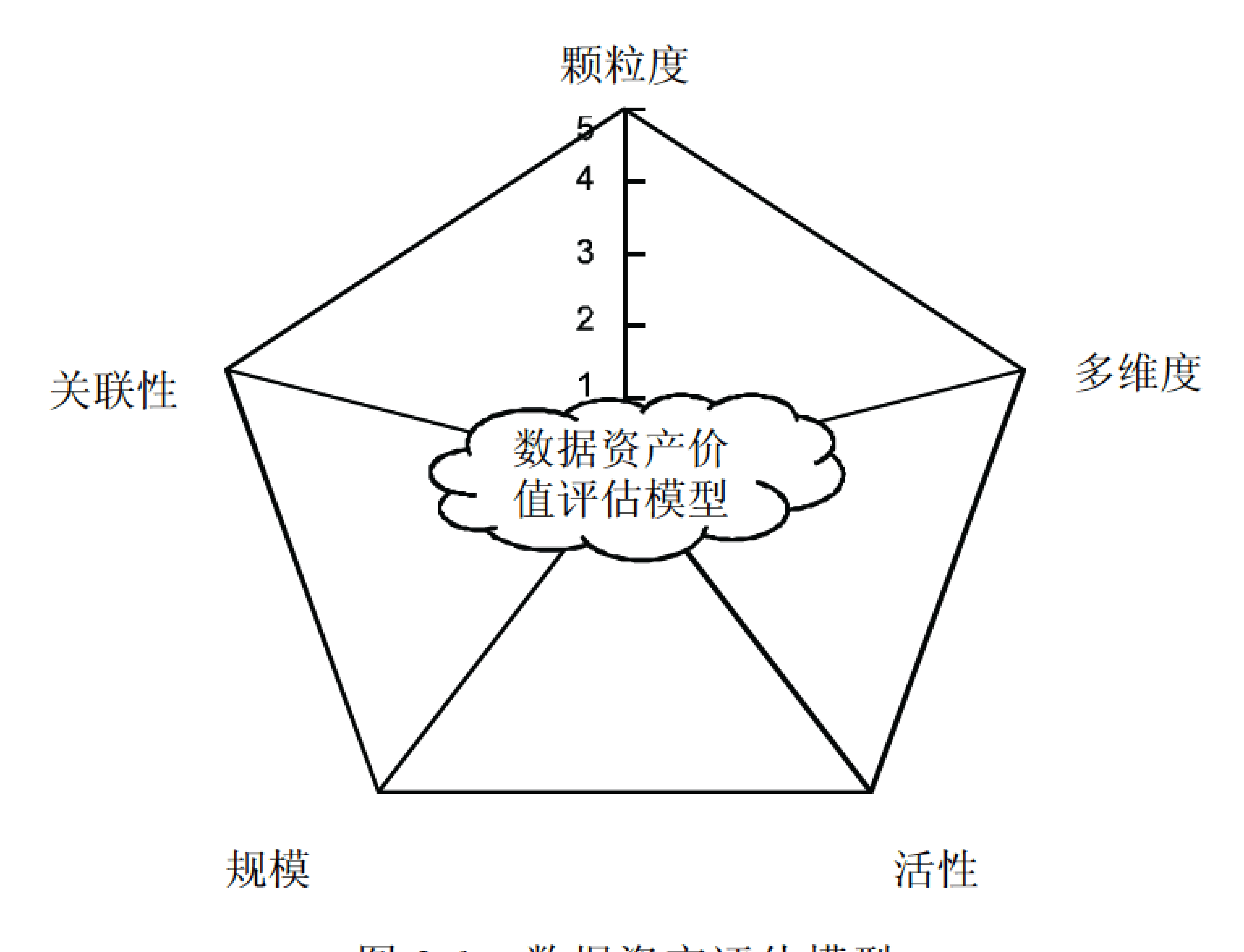


图 3-1　数据资产评估模型

所以把颗粒度作为反映数据资产质量的第一个维度。细化到一个人、一件单品、

一个网格、一个门牌号、一个零件，夸张地说，就算是一粒沙，也要清清爽爽地记录下它的位置、大小、重量，甚至因风吹浪打漂流的轨迹。不要忘了“一沙一世界，一花一天堂”。

多维度指标借用空间维度的概念，来指代数据来源的丰富性。每增加一个数据维度，会影响所有原数据的分析和判断，甚至会带来颠覆性的证据。

FICO 信用评分，是美国评估个人信用级别的通行标准，几乎每个美国人都有一个 FICO 评分。当人们申请信用卡、汽车贷款、住房贷款时，大多数的信贷机构都会参考申请者的 FICO 得分。但是在其发展的初期，FICO 模型中，仅仅依赖申请人在现有住址住了多久、为现在的企业工作了多长时间、申请人账号开设了多久等数据。根据这个评估标准，几乎所有 30 岁以下的人，都会存在很大的风险。现在人们知道淘宝上的购买主力，恰恰是以年轻人为主。所以零售商们群起反对，这些条款限制了发卡人数，不利于刺激消费。当 FICO 增加了评估数据的维度后，譬如纳入教育水平、职业等指标，那些受过良好的教育、从事体面职业的人，也就获得了信用卡。事实证明，他们的违约率极低。

在多维度指标中，人们尤其重视一类“先验”数据维度。譬如，人们在买股票的时候，一定先观察一支股票的行情走势；人们在买商品的时候，一定会对比和询价。互联网有助于把这些数据收集起来，进行分析，可以预测未来人们是否会买入股票或者商品。

在一次路演后，笔者和一家机构的投资总监聊大数据的发展前景。恰好这位投资总监熟悉 IT，而且还兼管公司旗下基金的发售。当他听到“先验”数据的概念，非常兴奋，反过来给笔者讲，如何利用这些数据来挽留老客户。他说：“如果购买了我公司基金产品的人，经常访问其他公司的基金网站，毫无疑问，这个客户换基金的可能性就会大大提高。我们必须在这个时候去干预，主动联系这个客户，搞清楚原因。”

活性指标的命名，带有感性的色彩。其原意是指生物体内发生的生理过程或处

于活动的状态或属性。数据的活性，指代数据被更新的频次。频次越高，活性越大。Facebook 公司在 2012 年 10 月，庆祝月度活跃用户超过 10 亿个。这里的活跃用户和数据的活性紧密相关。股民对换手率指标非常熟悉，换手率标志股票交易是否活跃，成为判断股价走势非常重要的指标。

曾经有一家公司过来咨询，他们的数据能否算作大数据。这家公司收集了大量的用户缴费数据、譬如交水电费、煤气费、有线电视费等。毫无疑问，这些数据非常有价值，但就是活性稍差，用户缴费最多也是一个月交一次费用。

新浪微博的数据，无疑是最具活性的数据之一，体现出实时的价值。利用微博数据，进行实时的精准营销，是许多公司孜孜以求的目标。

规模指标最容易理解。没有“量”的积累，就没有“质”的突破。数据量的增长，即是数据规模的扩大。但是到底有多大规模，才能算是“大”数据，的确是各行各业都很关心的问题。譬如互联网应用，如果没有 1000 万用户，估计很难称为大规模。但是如果一家券商拥有 1000 万个 A 股账户，那绝对是呼风唤雨的“老大”。规模这个指标很重要，但不需要执着于此指标。不同行业，不同的业务特征，对规模的定义完全不同。数据思维要先行于数据规模。

关联度指标反映不同多维数据之间的内在联系。之所以把关联度拿出来单独讨论，主要原因是同一企业内部存在大量的“孤岛”现象，不同部门之间积累的数据无法融合，形不成合力。造成这个现象的原因非常多，详加讨论超出了本书主题范围，在这里只是单列一个评估指标，提示数据融合的重要性。

富含“行为信息”的数据资产具备明显的“魔法水晶球”特征

在第一章笔者就提出，大数据令人着迷的一个鲜明特征就是对未来的预测。在此花些篇幅来详细的说明，哪些数据资产更加具备预测能力。

“物有本末，事有终始，知所先后，则近道矣。”《大学》中的这句话，其实讲的就是因果规律。如果知道某件事情发生，就会知道另外一件事情发生，这就接近了

解事物的本源了。古人说得非常富有韵律和诗意:“有道之士，贵以近知远，以今知古，以所见知所不见。故审堂下之阴，而知日月之行，阴阳之变；见瓶水之冰，而知天下之寒，鱼鳖之藏也；尝一脔肉，而知一镬之味，一鼎之调。古观其象，识其数，明其理焉。”

人们把 A 事件发生，一定导致 B 事件发生，称为强因果关系。更多的情况下，A 事件发生，可能导致 B 事件发生，反过来 B 事情发生，则 A 一定发生，把这种关系定义为“弱因果关系”。弱因果关系在生活中大量存在。如果人们要购物，则一定会询价，问问老板这件衣服、那辆汽车的价格，但是询价并一定导致购买行为。炒股票也是同样的道理，买股票之前一定会先查看行情，但是看行情，未必一定导致买入股票。数学好的读者，一定清楚，A 不过是 B 的“必要条件”。

互联网的普及，使得大量的 A 类型事件被自然而然地记录并数字化，云计算为这些数据提供了存储的空间，大数据则真正发挥这些 A 型事件的价值。人们可以统计大量的 A 事件和 B 事件，计算出“可能性因子 δ ”，利用“ δ ”乘以 A 事件发生的数量，就可以得出 B 事件的数量，获取精准预测未来的能力。

《南方周末》在 2012 年 7 月，刊发了一篇“德温特资本市场”公司的采访稿。这家公司的创始人是一位 80 后，名叫保罗·霍廷（Paul Hawtin）。他利用计算机程序，对全球最大的微博客推特（Twitter）上的推文进行抽样，抓取如“我感觉”、“我认为”、“……让我觉得”等表达投资者和公众情绪的语句进行分析、归纳，然后做出预测金融市场走势的判断，并声称预测成功率达到 87.6%。

这家公司未来的成长，还有待观察，但是这个事情理论是可行的。假如保罗·霍廷把数据来源换成雅虎财经、新浪财经等各大财经网站上所有用户浏览个股和大盘行情的数据，毫无疑问，的确可以预测股价的走势，甚至是个股的走势。

利用弱因果关系赚钱的方式多种多样。在信贷市场，公司的稳定经营是还款能力的必要条件。如果有大量的数据表明公司具备稳定、持续经营的能力，无疑会大大降低还款的风险。所以现在大型银行开始向一些大型网站购买这些数据，作为是否发放贷款的依据。

2012 年 6 月 28 日，建设银行“善融商务”电子商务平台开业。有一条宣传口号，“易商易融，买卖轻松”一语道破建行搞电子商务的天机，就是为了获得商家的经营数据，来降低其融资风险。建行没有打算在电子商务赚钱，而是把电子商务定位成助力建行信贷业务的工具。

不妨把这种有助于预测未来的“必要条件”数据，称为“水晶球”资产。事实上，凡是拥有内部网站的公司，都会拥有一类水晶球数据资产，这就是系统在运行时产生的日志。

假设有一家制造商，销售机构遍及全球，有什么好方法来管理各地分支销售机构的经营情况呢？

每周的例会，制定销售计划、跟踪销售漏斗（指潜在客户、购买意向、最终购买的数量都会梯次下降的情况）、分析成交数据等。这都是常规的手段，大家都是这么做的。但是这些管理手段都是后知后觉的，大部分是在事情发生之后，采取的控制手段。

做一个思想实验，会发现以下类型数据非常有效。销售人员在推销新的产品之前，都需要学习培训，包括讲课、自主学习等。客户在使用产品的时候，也会需要说明性的文档，这些文档往往都是销售人员提供的（针对企业销售的例子，并非针对个人消费品）。通过分析系统的日志发现，凡是销售业绩出众的分、子公司，有助于销售产品的电子文档，被下载访问的次数要高于其他分、子公司。这是一个很有意思的事情：总有一些“天才”销售员，可以不用借助常规的销售手段，获得出色的销售业绩。但是绝大多数的销售员，必须按部就班、刻苦学习、辛勤工作，才能有所收获。恰巧网站忠实地记录下了这些销售人员辛辛苦苦的点点滴滴：每天几点访问网站，下载资料，在“学习区”逗留多长的时间，每天访问多少次等。把这些数据汇集起来，形成销售员的“勤劳镜像”，再加上公司内部客户管理系统中记录的销售员拜访客户的时间、频次，甚至可以清晰地反映出每个销售员在拜访客户之前的准备工作是否充分等情况。

于是，这家公司开始深入分析系统日志，抽象出一些指标来度量销售员的勤奋

程度，进而可以形成分、子公司的勤奋程度。这些指标，居然和分、子公司的经营业绩紧密相关，而且往往提前一两个月就能反映出未来分、子公司的收入情况。这就为总公司及早干预，赢得了宝贵的时间。

这家公司的名字叫“华为”。专门提供系统日志分析工具的公司叫 Splunk，其上市第一天的市值就突破了 30 亿美元。

如果哪家公司的 CIO，能够像华为公司一样，充分利用公司内部的数据资产，形成对公司业务具体、细致的指导和预测，估计 CIO 就不再仅仅是技术官员了，而是公司决策层的核心之一了。

关于数据资产的几条建议

1. “天下武功，唯快不破”。越快地处理数据，越早地获取信息，就会越及时地做出商业选择。

2. 更多的数据来源，比更多的数据量更重要。这也是为什么数据资产评价模型中，把关联性和多维度作为重要的指标。

3. 数据富含多种信息，取决于观察视角。不要因为短期内没有用途，而随意丢弃。

4. 面对数据量指数般的增长，要早做打算。

5. 大数据不是核心问题，要聚焦于业务发展，善于从大数据中挖掘利于业务发展的信息。

6. 分享，而非保密。数据在流动中增值。流水不腐，户枢不蠹。

第二节　大数据飞轮效应是驱动产业融合的关键因素

提要：

1. 若公司能够利用客户数据为第三方开发出增值服务，就能支持公司持续地、免费地为客户提供更多的服务，而更多的服务产生更多的客户行为数据，利用这些数据又能为第三方提供新的增值服

务。这是个正向反馈的循环，如同巨大的飞轮，初始启动非常艰难、非常费力，需要持续不断地努力推动，才能有一点点效果，飞轮开始旋转很慢，但会越来越快，飞轮快速旋转时，只要一点点推动，就会产生巨大的效果。这就是大数据的飞轮效应。

2. 传统产业升级的典型案例：一家从事传统印刷业务的公司——雅昌，通过年复一年、日如一日的数据积累，成功启动大数据飞轮，轻松进入印刷、拍卖、艺术策划、展览、摄影、视频、电子书、艺术品收藏和艺术馆等多个行业。

那些拥有大量数据并且深谙大数据之道的公司，很快就会发现产业扩张的捷径。相反，那些对大数据感觉麻木、反应迟钝的公司，尽管一时风光无两，必将被新锐所淘汰。本节简要阐述大数据促使产业融合和分立的原理，详细内容和案例分析参见后续的章节。

产业融合指某产业不断侵蚀另一产业的空间，最终形成新型产业的过程；产业分立指因技术发展而成立的新产业，或者是附属产业成长壮大为新兴产业的过程。举例来说，我国十余年不见进展的“三网融合”，就属于通信业、互联网和广播电视产业融合范畴。加入大数据这个因素后，一些原本强大的行业，很可能变得不堪一击。顺便说一句，在产业加速融合的大背景下，分行业监管的措施和机构，已经成为严重阻碍产业升级的绊脚石。

大数据飞轮效应

考虑一种最简化的情况，公司销售产品给客户，客户付费给公司。在这种情况下，公司与客户之间的联系并不紧密，常见的方式包括在产品销售前的市场营销活动、销售拜访活动，钱款两清后，只要产品不出问题，基本可以不相往来。日用品、耐用消费品大部分属于这种情况。公司在运营中，难以采集有效的用户数据。信息系统中，最多会记录产品销售的批次、品类，难以获取具体消费者购买的某个单品

数据，如图 3-2 所示。（注：在本节中，并未严格区分客户和消费者的定义，为行文简洁，客户和消费者相互指代。）

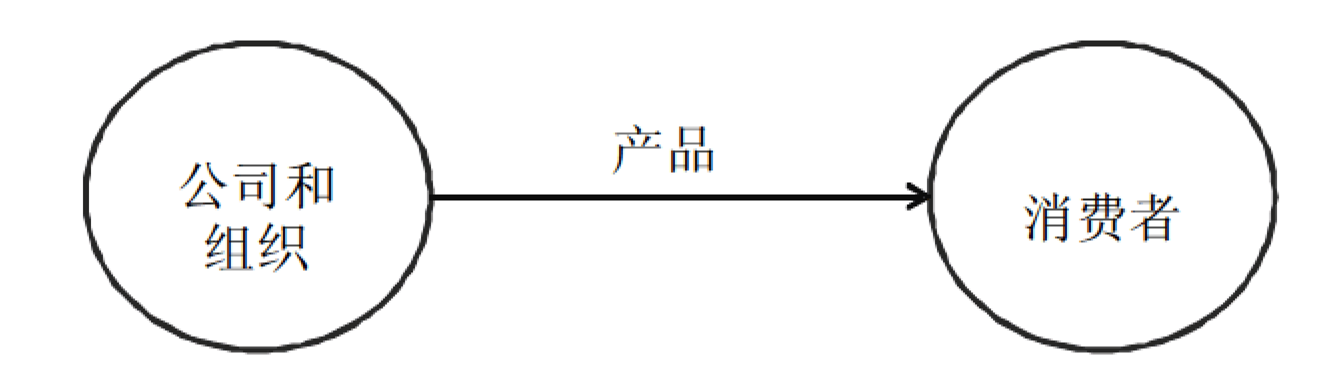


图 3-2　仅仅提供实物形态的产品，难以获取消费者的使用数据

第二种情况，则是公司给客户提供服务，客户付费享受服务，但是客户不得不提供给公司客户的基本资料，如图 3-3 所示。譬如，人们享受快递上门的服务，就必须提供住址；享受通话的服务，就不得不提供对方电话号码；去看病，就需要提供过往病史给医生参考；去贷款，就需要提供良好的信用记录。人们在享受各种服务时，不可避免地让出了部分个人信息。电信运营商、金融公司、医院等服务机构，在运营的过程中，势必积累大量的客户数据。如何发挥这些数据的价值，同时避免泄露客户的隐私，是这些机构面临的最大挑战之一。

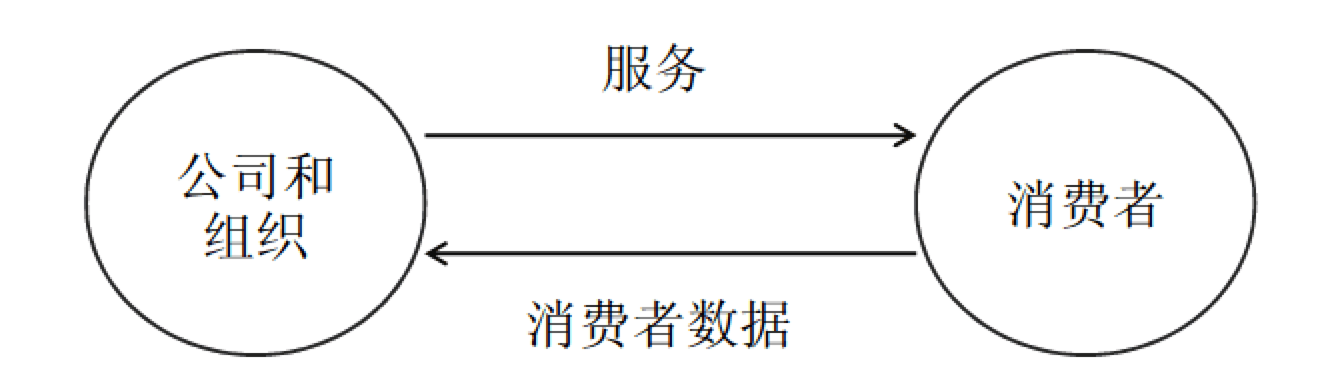


图 3-3　消费者享用服务的同时，不得不提供基本的数据，乃至使用数据

互联网是第三种服务模式。美国雅虎公司开创了免费服务的先河。雅虎通过免费的互联网址的分类服务，吸引人们把雅虎网站作为上网的首选地址。雅虎就像一条免费的高速公路，走的人越多，高速公路旁边的广告牌就越值钱。雅虎通过免费的服务吸引大量人流，通过收取其他公司的广告费来盈利，构成了一种三边的格局。这是多赢的局面，客户得到免费且高质量的服务；商家宣传了产品，增加了销售收

入；而雅虎公司赚取了广告利润。

这个阶段还算不上对某些行业的颠覆，充其量互联网行业分流了广告业的收入。在全球 5000 亿美元广告费支出的背景下，这点分流算不上什么。但在广告主眼中，互联网无非是另外一个展示商品的渠道。

互联网公司积累了大量的客户数据，尤其是在积累了大量的客户网上行为数据后，产业竞争格局发生了根本性的变化。凡是以信息传递为主的现代服务业，无不遭遇了灭顶之灾。这个名单可以列很长：通信服务、广播、电视、传媒、金融、物流、零售、中介、教育、医疗、文化以及公共服务。

从图 3-4 中就能看出，只要公司能够利用客户数据为第三方开发出增值服务，就能支持公司持续地、免费地为客户提供更多的服务，而更多的服务产生更多的客户行为数据，利用这些数据又能为第三方提供新的增值服务。这是个正向反馈的循环，如同巨大的飞轮，初始启动非常艰难、非常费力，需要持续不断地努力推动，才能有一点点效果，飞轮开始旋转很慢，但会越来越快，飞轮快速旋转时，只要一点点推动，就会产生巨大的效果。这就是大数据的飞轮效应。

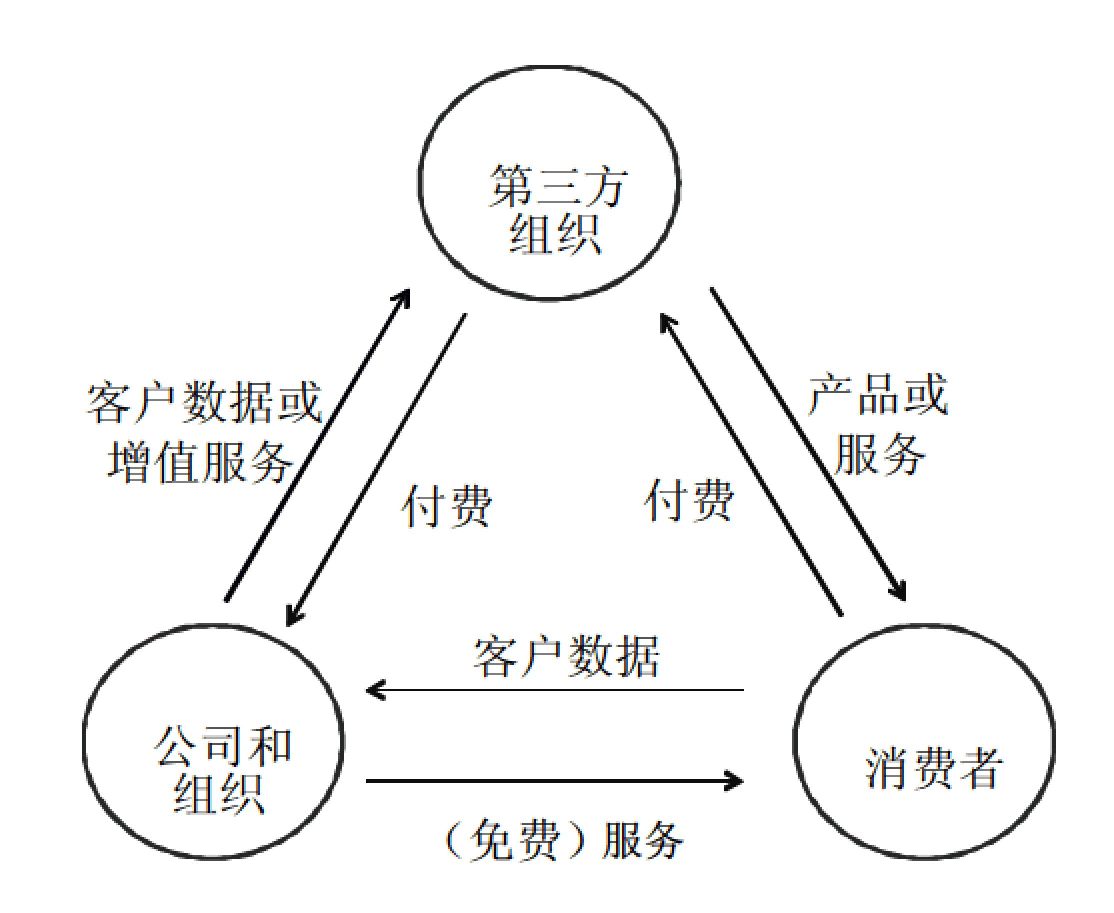


图 3-4　利用数据资产，可以帮助第三方改善服务

在大数据飞轮效应裹挟之下，公司从第三方获取的收入越多，就能够为客户提供更多的免费服务，尤其是基础性质的服务。不幸的是，语音通话服务，就被列入了基础服务。在大数据飞轮之下，此项服务行将免费。这些年语音通话资费不断下

降，就是强有力的证据。

大数据的飞轮效应还远不止于此。再看第四种情况，如图 3-5 所示，第三方依赖公司提供的客户行为数据，向客户提供产品或者服务。在这种模式下，如果公司中止向第三方提供客户数据，第三方就变成了聋子、瞎子。公司越俎代庖，可以直接向客户提供原本属于第三方领地的服务。笔者在这里做出一个大胆的推测，凡是依赖充分的客户行为数据分析开展的业务，都将被大数据飞轮所吞噬，谁率先启动大数据飞轮，谁就能利用飞轮效应获得产业竞争优势，在产业的融合与分立趋势中占据主动地位。产业界正在上演的一部大戏，就是互联网业利用大数据优势侵蚀传统的金融业。一些实力雄厚的金融巨擘，也摩拳擦掌进入互联网服务领域，就是此类大数据飞轮效应的完美注脚。

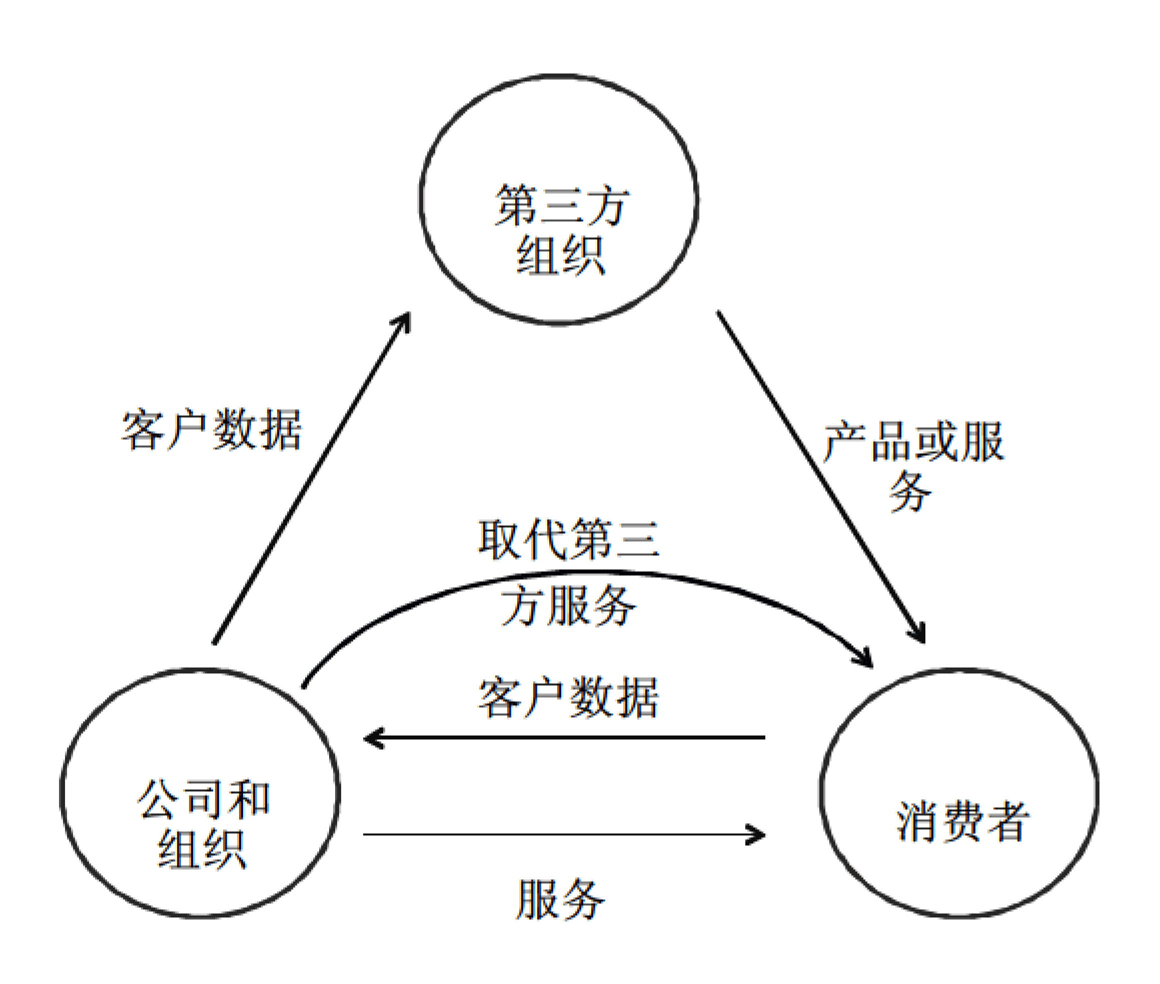


图 3-5　拥有垄断性的、独一无二的数据资产，则可以取代第三方的相关服务

第三节　一家“传统”公司的大数据飞轮战略

提要：

1. 为保存和传承雅昌数十年积累的宝贵数据资产“中国艺术品数据

库”，公司创始人立下企业遗嘱：“不管雅昌将来遇到什么困难，雅昌‘中国艺术品数据库’保存的数据不属于某一个人，也不属于某一个企业，它将永远属于国家、民族和整个人类。”

2. 在传统的印刷行业，雅昌实现了整个产业链的垂直整合。随后凭借着独一无二的数据资产，雅昌进入了印刷、拍卖、艺术策划、展览、摄影、视频、电子书、艺术品收藏和艺术馆等多个行业。可谓数据在手，产业随心。

雅昌是一家非常有趣的公司，一直致力于艺术品的印刷业务。北京申办奥运会、上海申办世博会、建国 60 周年大庆的画册等都是由雅昌印刷的，凡是代表国家抛头露面的印刷品几乎被雅昌包办了。大家有兴趣可以去雅昌艺术网了解详细的资料。这家公司与众不同之处在于：在精益求精地追求印刷质量的同时，一点一滴地积累印刷品电子数据，使公司从相对狭窄的印刷市场，成功转型为一家为大众提供艺术服务的新型互联网公司。它是印刷领域的垄断者，也是出版行业的新贵，其各类高仿真艺术品供不应求。雅昌在移动互联网领域亦如鱼得水，相继推出了“大千世界——张大千的艺术人生和艺术魅力”、“何家英师生展”、“王鑫生意象 · 蜕变”、“丽山园遗珍”等多款移动应用，其中“丽山园遗珍”获得了苹果公司教育类应用官方推荐。雅昌已经跨越了许多行业（见图 3-6），而未来它还能跨越哪些行业？

在搜索引擎中搜索“雅昌艺术网”，映入眼帘的是“雅昌艺术网——中国第一艺术门户网站”。打开网站，艺术品资讯、数据应有尽有，你能相信这是一家印刷公司开办的网站吗？可雅昌的确是印刷公司起家的，它充分利用拥有的数据，凭借对艺术品市场的理解，进军艺术品市场，并且在艺术品市场垂直整合。它是艺术品市场中的印刷企业，印刷市场中的艺术品企业。

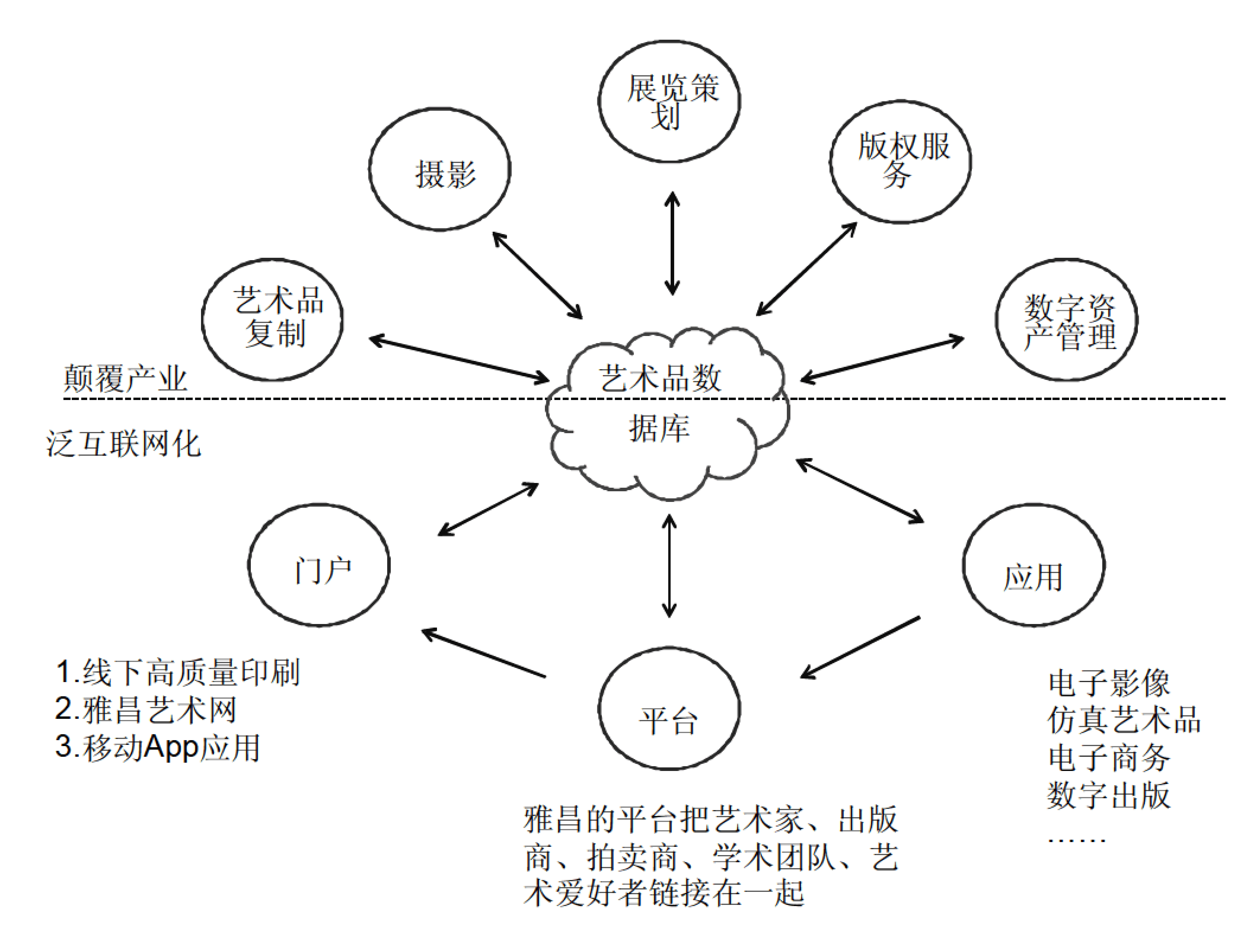


图 3-6 雅昌借助泛互联范式，启动大数据飞轮，开始跨行业扩张

1993 年雅昌公司正式成立，与其他同行不同的是，雅昌一直在坚持做一件看起来是“极大浪费”的事情——存储印刷的电子资料。在传统的印刷行业，这些资料往往被当作废料直接在电脑中删除。20 世纪 90 年代初期，一张空白的光盘就要卖十几块钱。雅昌长年累月地刻光盘，当时很少有人能够想象到这些堆积如山的光盘将是雅昌最宝贵的资产。截止到 2011 年年底，雅昌的数据中拥有 6 万余名艺术家、2000 多万件艺术品的电子图文资料，形成“中国艺术品数据库”。其创始人立下企业遗嘱：“不管雅昌将来遇到什么困难，雅昌‘中国艺术品数据库’保存的数据不属于某一个人，也不属于某一个企业，它将永远属于国家、民族和整个人类。”

如果没有这些珍贵的艺术品数据，雅昌将被一直禁锢在印刷产业。以“数据”开发利用为分界点，雅昌的发展历程可以分为以下两个阶段。

第一阶段：传统印刷行业

雅昌的最初定位只是印刷业，当时是拼技术和设备，其实也就是拼资金、拼胆

量、拼速度，为了发展和壮大，无论什么业务都接。但是传统的印刷业务竞争激烈，利润微薄，且各家的技术、设备差别不大。

艺术品印刷市场对印刷工艺的要求最为苛刻，印刷品的设计、制版、印刷、装帧都要达到顶尖的水平。雅昌聚焦在艺术品市场，是其和其他同行拉开差距的第一步。在这个市场，雅昌几乎形成了独家垄断的格局。印刷领域的“诺贝尔奖”——班尼小金人，雅昌收获了 19 尊。

第二阶段：雅昌艺术网上线为标志，正式开始了数据掘金之旅

积累数据：建立艺术品数据库

之前，雅昌作为艺术品印刷公司，虽然与艺术沾边，但始终位于价值链的底端，还是在印刷这个夕阳产业中。但就其资源来说，多年的积累使雅昌在书画、文物、拍卖、摄影领域保存了大量艺术家、艺术作品的相关资产数据。这些珍贵的数据资料在传统的印刷业中往往因忽视而被丢弃，但是雅昌看到了数据的价值所在，给传统的印刷业注入了信息化的活力，艺术品不同于新闻，时间的积累会给数据带来附加的价值。依靠印刷业务积累了丰富的资源，雅昌建立了网络中的中国艺术品数据库，希望为客户提供增值服务，从而建立一座艺术品印刷领域的“银行”——储存拍卖行中国艺术品的拍卖数据，比如图片资料、拍卖时间、拍卖地点、拍卖机构、拍卖成交价等信息。根据不同的客户，雅昌将数据库分成四大类别：艺术品拍卖市场数据库、艺术家及作品数据库、书画印鉴数据库、画谱收录及书画著录数据库。

拥有中国艺术品拍卖数据与艺术家资源及其完整的艺术作品数据，既是中国艺术品数据库的核心价值，也是雅昌得以开发新商业价值的关键所在。建立了数据库，雅昌便拥有了核心资源，把艺术品市场的主体，包括拍卖公司、画廊、投资者、艺术家、印刷出版公司、艺术媒体、投资咨询机构、保险公司等吸附在自建的平台上，成为产业链中的重要一员。雅昌从一个出版公司成为了一个数据拥有者。

2000 年 10 月，雅昌再将数据进行整理加工并在前台进行展示，在中国艺术品

数据库的基础上开设了一个艺术门户网站——雅昌艺术网，雅昌的行业边界再一次扩大到信息发布者。在雅昌艺术网上，可以看到艺术品预览信息、艺术界动态消息、拍卖资讯专题、拍卖品浏览统计报告、拍卖品现场直播、成交成果公布、中国艺术品行情等信息，同时也给很多展览做预展。这些信息为艺术品提供了营销的渠道和信息的服务。雅昌将数据库作为核心，以雅昌艺术网为平台，极大地扩展了行业的疆域，辐射效应巨大。比起以前做印刷的收入，雅昌艺术网的收入模式丰富了很多，包括收费拍卖品直播、广告收入等，这足以令还处在水深火热竞争中的印刷企业羡慕不已。

挖掘数据：雅昌艺术市场指数

艺术品市场中，艺术品拍卖市场份额最大，在 2011 年，艺术品拍卖的市场规模达到了 975 亿元，占艺术品行业规模的 46%。艺术品市场中，拍卖是站在艺术品金字塔顶端的位置，拍卖是艺术品市场的风向标，引领着艺术品市场的走向。雅昌虽然不参与拍卖，但是却是行业规则和标准的制定者。2005 年，雅昌开始发布艺术品拍卖行情，基于已有的庞大数据库，聘请证券公司研发人员进行几年的研究，推出了“雅昌艺术市场指数”，包括成分指数、分类指数、个人作品成交价格指数三大类艺术品市场指数。AMI 就像股票指数一样，成了艺术品投资分析工具和艺术品市场行情的“晴雨表”，开启了文化产业链的引擎。雅昌用艺术市场指数的方式将艺术家、拍卖公司、艺术品买家等紧密联系在一起，使他们对雅昌形成了强依赖关系，就像每天要看天气预报一样，将雅昌打造成为艺术市场的支柱。国内外许多买家、收藏家正是通过雅昌艺术网了解到艺术品的信息后参加竞拍的。数据如果不加以应用和挖掘，其价值仍然得不到体现，雅昌将已有的数据进行了价值的二次开发并且通过网络信息的快速扩散做大了艺术品行业的蛋糕，自身也迈向了价值链的上游甚至顶端。

以数据库和信息技术为基础，垂直整合产业链

在拍卖市场之外，雅昌利用自己的信息化优势，基于数据库的资源在产业链的各环节为各个主体提供增值服务，包括数据存储、整理以及用各种方式呈现的服务。

首先是数据存储服务和数据的印刷和展示服务。对于艺术家，雅昌提供数字资产管理、官方网站、艺术家文献库、出版等一站式服务，以提升艺术家的核心价值。2012 年 3 月，中国美术家协会副主席冯远使用了数字资产管理服务。雅昌通过授权，将艺术家作品底片电分为标准 CMYK 四色数字文件，同时将他们在雅昌印刷作品的高分辨率图片一并存储在数字资产库中，相当于为艺术家在雅昌建设了个人专属的作品数字档案馆，这些数据资料可以为艺术家建设个人官方网站、艺术衍生品设计制作、版权代理、出版个人画册等所用。

雅昌艺术网作为网络平台，不仅极大地消除了艺术品行业的信息不对称，也打破了地域、收入等限制，使艺术品市场走向千家万户，同时其自身也开拓了业务范围。

艺术品市场是个小众市场。传统的艺术品市场门槛高、鱼龙混杂，没有专业知识的人一般涉足较少，因此也掩盖了人们对于艺术品的需求。但是雅昌艺术网提供了低门槛、低成本的途径，网站用户可以搜索、复制自己喜欢的艺术品。在高端印刷的基础上雅昌拥有了卓越的复制技术，便提供了高端艺术品复制业务，一些名家会授权限量复制其作品，雅昌的高端艺术品复制使珍贵的艺术品再现后成为生活必需品，雅昌随之走进了艺术衍生品市场。同时，雅昌也构建了人们网上购买艺术品的渠道，运营了雅昌交易网，为艺术品、收藏品和艺术衍生品的买卖双方提供交易服务平台。

在存储服务的基础上，雅昌推出了雅昌“艺+”作品认证服务，将艺术家的作品进行收集，再次纳入自有数据库，而且将存入的数据做了标准化处理，更有战略意义的是，这种数据成为了行业的标准和认证的权威，这种滚雪球式的积累和应用，转起了雅昌大数据的飞轮。

雅昌“艺术作品”认证是通过对艺术家作品的数字化图片进行真伪认证，为每一件真迹的数字作品提供永久性的唯一编码和标识。经过认证的艺术作品文件将会保存在雅昌艺术家个人数据库的“中国艺术品认证系统”中，并通过其官方网站（www.fengyuan.artron.net）公布。收藏家、投资人、艺术爱好者等可在线查询，使收藏者能够及时甄别作品真伪，杜绝仿品流通，更为维护艺术家的个人品牌以及知识产权提供了有力的保障。待数据库足够丰富后，雅昌拥有了完整的中国艺术品真迹的数字版，也就是网上故宫。在不久的将来，数据金矿在雅昌的战略下还将在商业模式上大有作为。

具有了数据，必然就会有搜索的需求，雅昌就此推出了“雅昌艺搜”的垂直搜索产品，雅昌帮助客户在这个信息非常少的垂直领域找到需要的信息。

与线上数据资料收集的思路相一致的是，雅昌在与拍卖行、画商和画家的长期接触中，有计划地进行艺术收藏，与大师结成合作伙伴。2006 年 5 月，占地 1 万平方米的雅昌艺术馆在深圳开馆，作艺术品收藏、展览之用。凭借着与大师的伙伴关系和已有的资源，雅昌帮助大师们策划各种展览和学术讨论活动。

如今世界艺术品市场已经逐渐成熟和规范并且呈现出阶梯状的三级市场，分别为基础产业画廊业、艺术博览会、艺术拍卖会，雅昌已经全面布局，在艺术品行业气贯长虹。

走向移动互联网：雅昌电子图录

2011 年春季的拍卖市场上，很多竞拍者都是拿着 iPad 来参加拍卖会的，因为他们都在使用着“雅昌拍卖电子图录”应用。它基于雅昌权威的拍卖数据，为收藏家提供完整而精确的拍品信息、市场价值走向、相关艺术家分析等重要资讯。iPad 用户可通过 AppStore、Google Store 等软件商店下载客户端专用软件，进而下载用户所需要的拍卖图录。用户可随时随地进行线下阅读，在看到喜欢的拍品时，还

可以随时收藏拍品、输入批注信息，或者加入电子书签，方便下次阅读查找。用户可以通过 PC、MAC 等终端，登录雅昌艺术网，下载任意一场拍卖会的电子图录进行浏览。

雅昌作为一家传统企业，凭借着对信息技术和数据的卓识远见，将艺术品数据库、雅昌艺术网和线下的艺术馆有机整合，将“IT+艺术+印刷品”的商业模式演绎得淋漓尽致，这是战略上的成功，打破了行业边界。雅昌通过权威拍卖指数、雅昌艺术网等树立了行业地位，并吸引到众多客户，为客户提供印刷业务或数据业务的同时，再次收集和积累数据及资料，越发庞大的数据库可以向客户提供更多的增值业务和其他印刷产品。对于上游的艺术家，雅昌利用线上的信息系统和可存储等优势提供数字资产管理、建立个人网站等服务，线下则提供印刷、策划展览等服务；对于中游的拍卖公司，雅昌网上提供预展，同时丰富数据库，线下进行拍品目录的印刷；对于下游的艺术衍生品，雅昌拥有先进的技术和数据资料，可以进行大量的商业开发，出售大量衍生品，包括图片影像、艺术品复制等。雅昌将信息的整合、艺术品的整理、客户需求的获取和理解有效地结合在一起，是跨行业经营的典范。

雅昌业务轨迹回顾

印刷—数据库—艺术门户网站—艺术品行情发布—拍卖—持续积累和垄断艺术品数据—数码艺术资产管理—艺术策划、展览、摄影—衍生品：CD-ROM、视频、电子书—艺术品收藏和艺术馆。

大数据的魅力是无穷的，它拥有的魔力像一盏灯，能够帮助一家企业突破原有的行业疆域和边界，向行业以外扩张。大数据照亮的范围都是企业在娴熟地利用信息和数据之后所能占领的领域，可以离原有行业很近，也可以很远，关键在于企业对数据和商业模式的理解。企业拥有了广泛的产业数据，不仅拥有了对产业基本信息的理解和洞察，更珍贵的是拥有了别人没有的生产资料，数据这种生产资料能够直接衍生出商业价值。那么拥有了产业数据的企业，便是产业的主宰者和规则制定者。

大数据时代，传统的产业概念需要全新的审视。很多时候，只要拥有足够的数据资产，“所谓产业，由你随便划”。

第四节　以数据资产为核心的商业模式

不同的产业，不同的数据资产利用方式，可以衍生出不同的商业模式，如图 3-7 所示。在数据加工的产业链上，依次有采集、整理、分析、决策支持环节，每个环节都是培育众多公司甚至明星公司的土壤。从产业维度来看，媒体、零售、金融服务、制造、医疗等行业，都需要嫁接大数据技术，或者提升主业，或者跨领域、跨行业扩张。计算机技术提供商，藉由大数据机遇，可以更全面、更深入地介入到客户的业务和生产运营中去。产业融合的大幕，正徐徐拉开。本节约略盘点部分公司，它们是各类商业模式的代表。

商业模式的思考角度，是站在企业外部观察企业盈利来源的变化而得出的。第一种商业模式是简单的出售或者出租数据。这些数据是精心采集而来的，公司保证这些数据的完整性、准确性和真实性。这个商业模式覆盖大数据产业链中最上游的两个环节，数据采集和整理环节。第二种商业模式在广泛采集的数据的基础上，去芜存菁，提取更有价值的信息，然后把这些信息打包卖给客户，称为租售信息模式，它覆盖产业链的采集、整理环节，部分工作涉及数据分析环节。

谈这两种商业模式，必然涉及到数据、信息概念，前面的章节并没有花费笔墨去解释什么是数据，什么是信息。再延伸一点，知识、智慧又和数据、信息有什么内在联系呢？笔者翻看各百科书给出的定义，大多比较拗口、晦涩，而且篇幅也较长。早些时候，笔者曾经发了一条微博，被华为公司刊物引用：“数据就像海边的沙；信息是埋在沙里的珍珠；历经淘洗反复筛选，制成的项链就是知识；能把精美的项链戴在心仪姑娘胸前，是为智慧。”

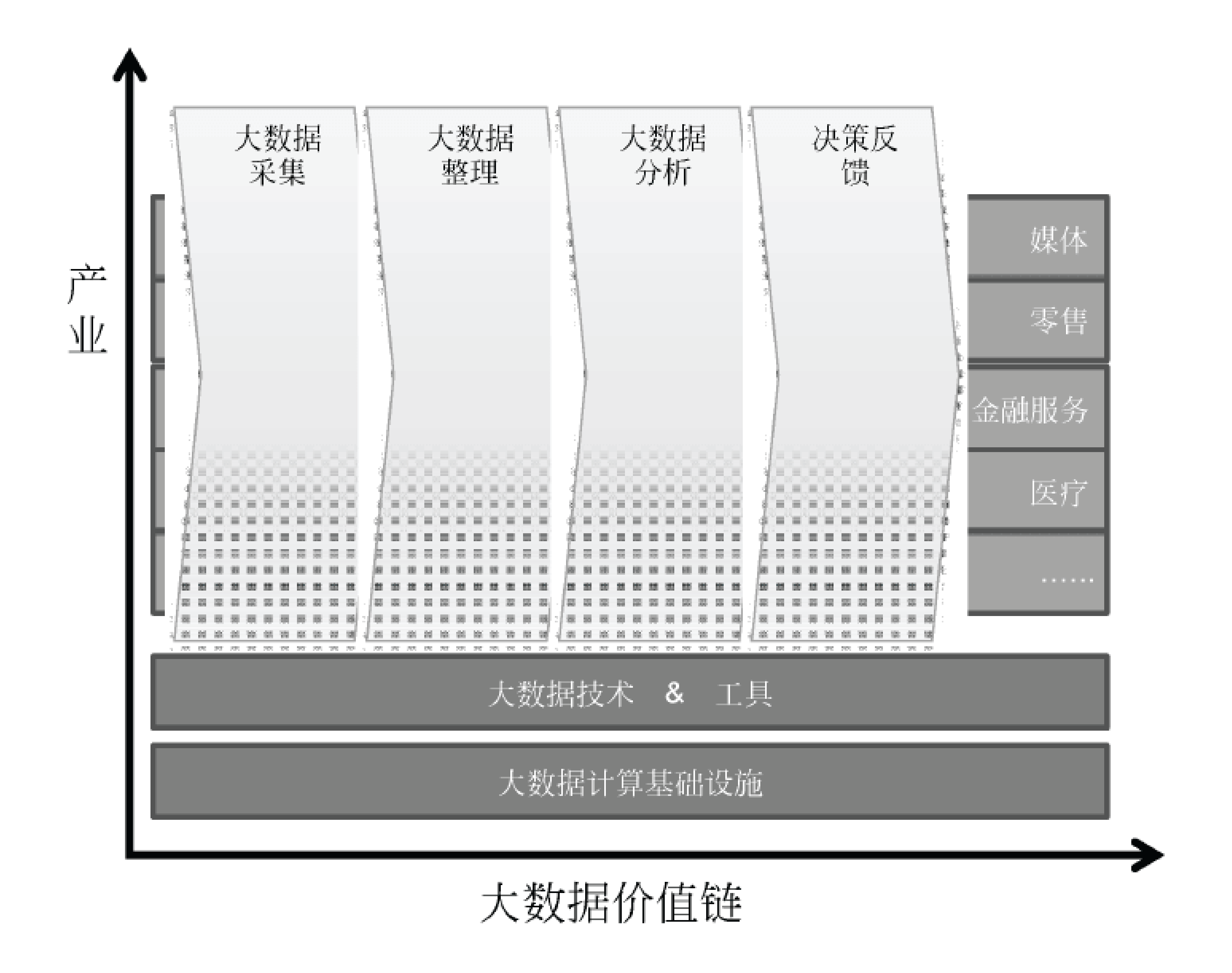


图 3-7 大数据价值链与商业模式[①]

沿这两种商业模式引申下来，大家自然而然会想到有没有售卖知识的商业模式？咨询产业属于这个范畴。如果咨询能力对接大数据分析、挖掘能力，将会引发大数据产业链顶端“决策”层的变化，公司战略决策模式甚至为之改观。但目前尚未发展到这个阶段。传统的商业智能分析工具如果没有大数据的支持，能够发挥的价值会有很大局限性，无法深度影响公司的战略决策。这也是笔者没有把传统的商业智能公司看成是大数据公司的主要原因。

大数据技术和产业对接，将迸发出前所未有的巨大商业潜力。笔者把缺少数据支持，而难以开展甚至无从开展的业务，定义为“数据使能”的商业模式。“数据使能”实际上是大数据对传统行业的颠覆和冲击。因为数据这个要素，使原本井水不犯河水的不同行业的边界日趋模糊，占据数据资产优势的行业，会不断蚕食、侵袭、颠覆那些处于数据资产劣势的行业。

谷歌在数字媒体领域攻城拔寨，居然会威胁电信运营商的生存空间！广告与传

① 图 3-7 是网友@尹锴_ink 看到笔者的博文《大数据时代的三大发展趋势和投资方向》后，帮助整理的，特此感谢。

媒产业，是另一个巨量的市场。根据网络营销咨询公司 eMarketer 的报告，全球各类广告支出逾 5000 亿美元。大数据技术推动的精准广告、计算广告份额不断地侵蚀传统平面媒体的市场份额。笔者把融合大数据技术的传媒公司盈利模式，称为“数字媒体”型。大名鼎鼎的谷歌、百度被人们戏称“外事不决问谷歌，内事不决问百度”，就是这种商业模式的典型代表。目前这个市场依然处在高速的变化之中，数字媒体商业模式将在第四章专门论述。

譬如金融业如果充分利用大数据，理论上其客户群可以扩大到近 5000 万家中小企业。但是非常遗憾，互联网公司尤其是电子商务公司，积累了大量的中小企业经营数据，它们更容易侵入银行腹地，开始为小型企业提供贷款服务、供应链融资服务等。相反，传统银行苦于缺少这些宝贵的数据资产，开展类似业务时陷于弱势地位。互联网公司携大数据优势，却不断地攻城拔寨。金融产业和大数据的对接是一个庞大的主题，将在第五章专门论述。

因为数据的重要性，围绕个人、企业数据资产的争夺，也成为硝烟弥漫的战场。为个人、企业提供在线网络存储服务的商业模式，称为“数据空间运营”模式。这种模式最早出现在国金证券大数据系列报告第三篇中，报告发布后，谷歌、联想、百度、京东商城等各界巨头纷纷推出类似业务，譬如新浪网盘、华为网盘等。现在媒体给出一个新词“个人云”，指代各类网盘服务，并纷纷预测 2013 年是“个人云”的爆发年。这种商业模式非常重要，但是国内市场独立运营网络存储服务的公司，日子恐怕不太好过。

大数据技术提供商，在促进大数据在各行各业的落地上功劳卓著。不同的数据类型需要不同的处理手段，尤其是非结构化数据，处理语音的技术拿到图像领域未必适合。因此，每一类非结构化数据都可能催生出一家大型的技术公司。它们如果凭借技术优势，在某个业务领域获得突破，将是非常值得关注的投资对象。对这些公司的介绍，笔者放到本书的第三部分。

租售数据模式

这种模式比较容易理解，和销售普通的商品没有太大的区别。但是，广泛调研发现，有的公司仅仅通过卖数据可以获得数亿的销售收入，有些则勉强度日。细究之下，营业收入的高低，与数据采集的技术含量、垄断性和销售渠道的垄断性紧密相关。

有人发了这样一条微博："……#数据推荐#【京东上的用户评论数据】数据集包含京东上 31 万用户对 18000 件商品的 165 万条用户评论数据。京东数据的好处是对正面评论和负面评论都做了分类，可用于评论分析、情感计算、用户行为分析等研究领域。数据大小：93.46M ；数据来源：自行抓取；数据详情：……"利用爬虫技术（一种自动获取网页内容的程序），可以实时获取电子商务网站上各类产品的价格。有些生产商以此来作为生产的依据，于是雇佣人手，专门监控某一品类的价格变动情况。有些嗅觉灵敏的技术高手，就开始提供这方面的数据抓取服务。

普通的爬虫技术可以从开源网站获得，技术门槛比较低。从业者虽人数众多，但绝大多数的规模都很小。另外，新闻中经常出现盗卖用户资料的事情，游走在法律的空白地带，绝大部分涉嫌侵犯公民的隐私。这种做法虽然可以反映数据的价值，但不是这种商业模式的主流。

真正通过租售数据获利，取得上亿营收规模，要么在数据的完整性、及时性、颗粒度上下功夫，要么在数据推送渠道上做文章。前者的代表是四维图新和高德软件，后者的代表是广联达。

案例：四维图新——销售地图数据

地图一直是国防、生活的必备品。红军长征时，刚刚进入云南，就是因为道路不熟，困于山区不得转移，恰好在缴获的军阀物资中发现了云南省地图，真是绝处逢生，如获至宝。我国制图历史也源远流长，在晋代制图学家裴秀已经总结提出了

绘制地图的六条原则——绘图六体。21 世纪，地图跟大众的生活息息相关。汽车导航、智能终端的普及，导致对精确地图数据的旺盛需求。四维图新就是致力于提供电子地图的公司之一，其营业收入的规模接近 10 亿元人民币。

四维图新号称拥有全国最大的高质量导航电子地图数据库，建成了以北京为中心、覆盖全国的本地化导航电子地图数据采集更新体系，在基于静态的地图数据基础上不断地加入实时动态的交通信息、丰富的生活信息和全面的地理信息，并通过了国际顶级汽车集团的全球采购质量评价，获得了国际上几乎所有主流车厂的订单。

在交通信息服务领域，四维图新形成了覆盖北上广深等 20 余个主要城市的服务网络，开发出了动态交通信息采集、处理、发布技术，目前已开通 22 个城市的商业化服务，应用于车载导航、便携导航、互联网、移动位置服务等领域。

四维图新在行人导航地图数据中包含了北京、上海、广州、深圳、香港、澳门等 20 余个城市的公交换乘、地下通道、过街天桥、人行横道及更多行人设施的信息，利用这些细颗粒度的数据，可以实现由车到人的全程导航服务。

四维图新毫无疑问地在地图数据采集、制作方面的能力非常强，而且数据的颗粒度也非常细，甚至每一个桥洞都被详细的定位，但是数据销售渠道上过渡依赖汽车制造商、手机制造商的预装，如图 3-8 所示。相比之下，广联达公司数据采集、制作的难度要低于四维图新，但取得了令人惊讶的增长。原因之一，就是对销售渠道和数据使用终端的完全垄断。

裴秀“制图六体”

一为“分率”，用以反映面积、长宽之比例，即今之比例尺；二为“准望”，用以确定地貌、地物彼此间的相互方位关系；三为“道里”，用以确定两地之间道路的距离；四为“高下”，即相对高程；五为“方邪”，即地面坡度的起伏；六为“迂直”，即实地高低起伏与图上距离的换算。

地图如果只有图形而没有分率，就无法进行实地和图上距离的比较和量测；如果按比例尺绘图，不考虑准望，那么在这一处的地图精度还可以，在其他地方就会有偏差；有了方位而无道里，就不知图上各居民地之间的远近，就如山海阻隔不能相通；有了距离而不测高下，不知山的坡度大小，则径路之数必与远近之实相违，地图同样精度不高，不能应用。这六条原则的综合运用正确地解决了地图比例尺、方位、距离及其改化问题。

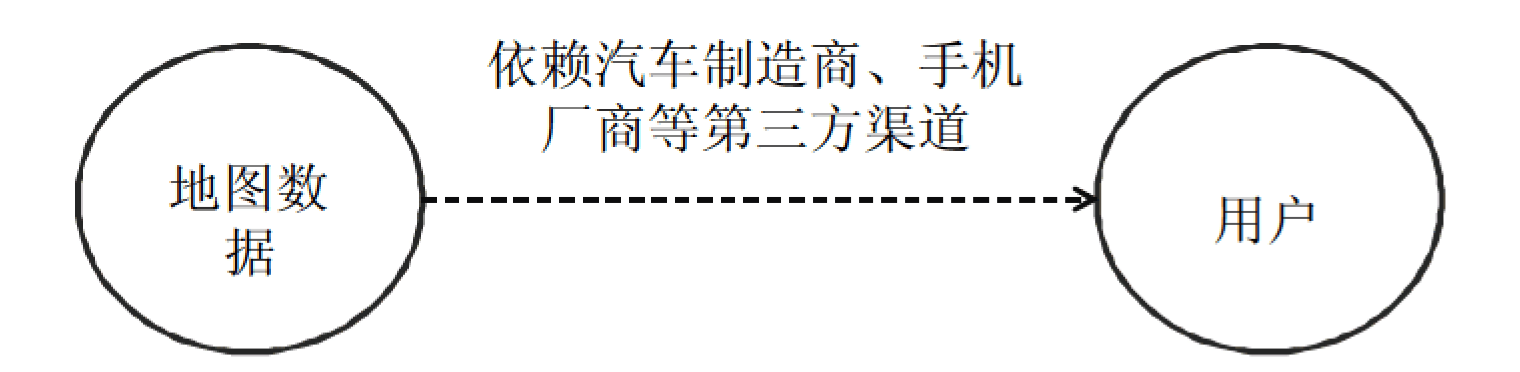


图 3-8　直接销售数据的商业模式，四维图新的数据依赖第三方渠道才到达最终用户手中

案例：广联达——从顺手牵羊到反客为主

广联达软件股份有限公司成立于 1998 年，是国内建设工程领域信息化服务产业的领军企业。2010 年 5 月，广联达在深圳中小企业板成功上市成为建设工程领域信息化产业首家上市的软件公司。广联达产品从单一的预算软件发展到工程造价、工程施工、企业管理、工程采购、工程教育、电子政务与互联网七大类 30 余种，并被广泛使用于房屋建筑、工业与基础设施三大行业，在建设方、设计院、中介公司、建材厂商、施工单位、物业公司、专业院校及政府部门八类客户中得到不同程度的应用。举世瞩目的奥运鸟巢、国家大剧院、上海东方明珠、广州东塔西塔等工程中，广联达产品均得到深入应用。

先看一段广联达数据服务的一份官方说明："广材信息服务是广联达公司推出的目前国内最领先的建材价格信息服务，在北京、上海、重庆、广东、新疆、南京、西安、武汉等十个重点城市设有上百人名专业的信息采集团队，为政府部门、开发商、工程造价咨询公司、设计公司、施工单位提供材料价格信息服务，其定期更新的数据库覆盖了建材及设备行业超过 500 万条信息及近 4 万家建材生产和销售企业。其服务包含'广材网'、'广材数据包'、'广材信息杂志'、'广材助手'等。广材网在线提供全国 300 万条材料的品牌最新市场价、数万家厂商联系信息查询、常用材料每日市场价格及行情走势、发布询价材料等专业材料信息。广材数据包提供覆盖土建、装饰、安装、给排水、通风、人工等专业的数据包。每条市场材料都和定额材料匹配，每条定额材料都有信息价、综合价、品牌价参考……"

建筑从业者看到这些说明，会一目了然，认可这些数据服务。事实上，广联达

仅仅通过销售建材价格数据包，收入就超过了 1 亿元人民币，成为公司的明星业务。这个业务未来的发展空间值得期待，很可能成为带动公司向服务、运营模式转型的排头兵。

简单梳理一下广联达业务的演进，有助于读者理解卖数据业务模式的特征。

广联达最初销售专业领域的工具软件，如工程算量软件、造价软件等等，并初步在同类工具软件市场形成了垄断地位。用户在使用这些软件时，为了获得更精确的计算结果，需要实时的建筑材料价格信息。这些价格信息是零散的，用户只能根据大概的数据来估算。广联达发现提供精确的建材价格信息，有助于用户更好地使用算量、造价等软件，开始为客户提供此类服务。当工具软件的功能和市场占有率达到一定程度的时候，建材数据反而成了用户最为关心的、最核心的问题，工具软件就蜕变成销售数据的渠道，数据反客为主，登堂入室，成为主角，如图 3-9 所示。

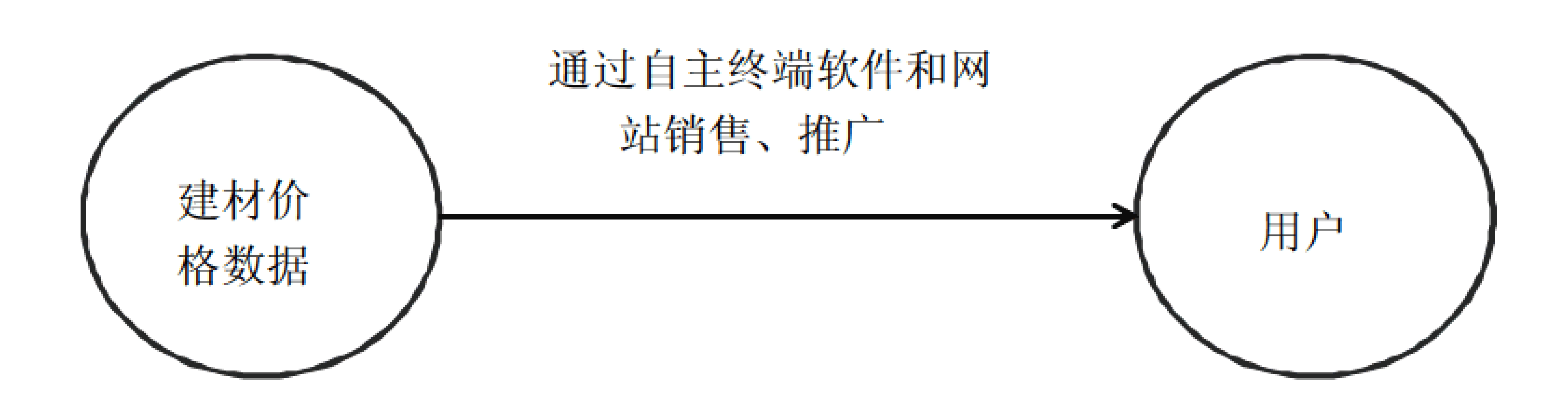


图 3-9 广联达的数据销售业务，建立了完整的渠道

租售信息模式

租售信息的业务模式和租售数据模式类似，将更加强调数据的广泛性、精确性、及时性，同时附加更多的行业特性，并且需要提供具备从多个角度、多种维度解读数据的工具。这种模式的典型代表是彭博、路透等金融信息服务公司。

华尔街的从业人员，上班第一件事情就是打开彭博的终端机，查看每日行情、各地政经要闻等等，甚至某个重大的事件，一定要等彭博刊登后才算确认。可见，彭博的信息在金融界的重要性和影响力。

彭博（Bloomberg）是现任纽约市市长创立的公司，致力于为金融从业人士提供及时、准确、丰富的金融交易信息和财经资讯。由于是非上市公司，笔者获取信

息来源主要依赖公开报道。公司核心竞争力在于积累了丰富、大量的金融行业数据和交易数据，通过彭博终端、杂志、电视、广播、App 等多种方式及时传递给用户极富价值的信息。在财经信息领域，有着 100 多年历史的路透集团和道琼斯曾是当仁不让的双寡头，但彭博的出现彻底改变了这个局面。现任纽约市市长迈克尔· 布隆伯格在 1981 年创建这家公司后推出彭博终端，这种双屏的终端在统一的平台上整合了新闻、数据、分析工具、报告和交易系统等多种功能，迅速占领了金融专业人士的高端市场。凭借收取高额终端费用和服务费用，彭博在 20 多年后成为全球最大的财经信息服务商，甚至逼得道琼斯退出了实时财经终端市场。按照彭博收集和监测的市场信息来看，其目前在全球财经信息服务市场的份额大概占 40%上下。

作为让它后来居上的法宝，终端机是彭博最大的收入来源。2010 年 11 月左右彭博终端在全球拥有 29.5 万用户，按照每个用户每月 1590 美元的收费计算，就意味着每年超过 57 亿美元的收入，这大概占到彭博总收入的 85%。

彭博公司主要提供三种业务类型服务：一是终端业务，在系统上彭博提供了 3 万多种功能，通过彭博的终端给资本市场的用户提供很多的深度报道、新闻和数据信息；二是交易系统的业务，交易系统的用户包括中国大的银行和大的基金投资者；三是数据业务。

为了完成上述服务，彭博拥有全球庞大的组织架构。彭博资讯在全球拥有 142 家新闻分社和 1650 名记者，还拥有全球唯一每天 24 小时播放财经信息的彭博电视台、彭博电台，以及在北美出版的 4 本专业杂志。由于功能众多，使用费用也相对昂贵。每个用户月租费为 1600 美元左右，年费合计将近 2 万美元，彭博终端产品在全球保持统一价格。

彭博的中文服务主要内容包括：实时中文财经新闻报道服务，中文财经资讯；中国的用户在彭博终端上还可以获得国内外的海量数据信息和搜索引擎；实时查询中国的公司债和国债价格及收益率曲线；实时查询中国汇市、货币市场及利率互换市场，并使用中国利率互换市场的计算器及定价工具。在中文服务推出之后，彭博

的服务对象和收费价格并未改变。

许多行业都存在类似彭博的商业机会。在中国金融信息服务市场，玩家众多，各擅胜场（见表 3-2），没有一家取得彭博一样的优势地位。估计这个市场存在大量洗牌的机会。

表 3-2　金融资讯服务商

类　　别	服务商代表
机构类销售产品	中诚信资讯、万得、新华、巨灵、港澳资讯、聚源、财汇、国泰安等
个人类终端产品	大智慧、同花顺、金融界、指南针、钱龙、通达信、东方财富等
私募类数据库	清科、朝阳永续、私募排排网、壹私募等
港股类	财华、多元世纪、经济通、新鸿基等
固定收益类	红顶证券、北方之星等
理财类数据库	普益财富、银率网等
数据服务商	华通人、塔塔、中经网、我的钢铁、化工网、煤炭信息网、布瑞克等
交易软件类	恒生电子、天软科技、携宁、盈时胜、金证等
研究报告类	银联信、今日投资、Microbell、启明星、维赛特等
财经网站	和讯、新浪、搜狐、腾讯财经频道

延伸阅读

对比四维图新、广联达、彭博的业务发展，发现建立端到端的数据销售渠道，掌控数据采集到销售的完整产业链，是销售数据（信息）商业模式所必须考虑的一项重大战略，如图 3-10 所示。四维图新的竞争对手，高德软件通过提供手机版导航软件，已经开始抢占最终用户，四维图新的战略发展方向值得关注。广联达一定会继续加强在工具软件方面的占有率，增加不同软件之间的协同，让数据在不同的工具软件间自由流动。

销售数据是四维图新的核心业务；广联达公司通过销售数据，实现和其他业务的“加和协同”效应；彭博则实现了信息的“乘数效应”，利用各种渠道扩大信息的

覆盖面。上一节里提到的雅昌，在商业模式上和彭博是相通的。它们都凭借独一无二的“数据资产”涉足多个行业，雅昌形象地称之为“一鱼多吃”的战略。

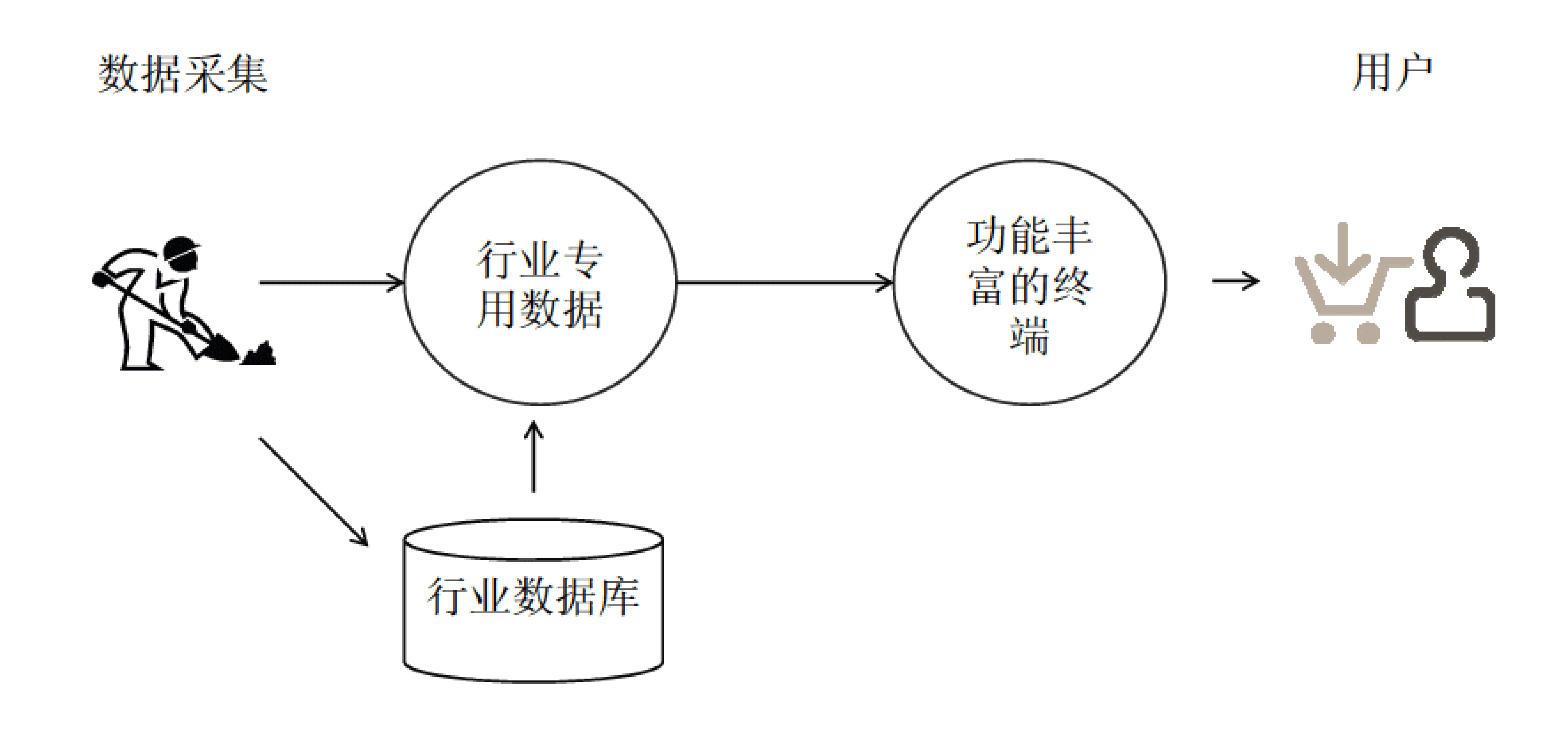


图 3-10 必须在终端领域有控制权，才能无忧运作销售数据（信息）的商业模式

行文至此，再来回顾第一章提出的泛互联范式，将更深刻地体会到其精微之处，产业的兴衰交替、公司的此消彼长，莫不与之相关。

数据空间出租模式

2012 年 6 月前后，百度网盘、新浪微盘纷纷升级到 100GB 的免费存储空间。京东商城也悄无声息地推出网盘服务，用于注册用户保存在京东购买的电子书等产品，也可以保存个人的文档。腾讯早已通过文件中转站提供临时的文件存储服务，现在也有网络硬盘，用 QQ 账号即可登录。百度百科上列出的国内的网盘提供商有 40 多家，海外的互联网巨头无一缺席，纷纷开展类似的服务。互联网公司推出网盘服务，无可厚非，但联想、华为等传统设备提供商，也开始进军网盘市场，意味着什么呢？

如果联想到“拥有数据的规模、活性”这句话，那么这种模式可不仅仅是租了一个虚拟空间，保存几个不常用的文件那么简单了。它的演进逻辑是从简单的文件

存储，逐步扩展到数据聚合平台。谷歌的商业模式，完全可能从空间存储领域重演。提供网盘服务的公司完全可能利用其中大量的数据，开发增值服务。提供精准的广告，只是当下流行的赚钱之道之一。譬如网易，为其用户提供的照片打印业务，就很有趣。网易用户可以选择自己相册中满意的照片，由网易帮助其把照片印刷到饮水杯或 T 恤衫上，然后给用户快递到家。数据自有黄金屋，数据自有颜如玉，先把用户的各种数据都抢到自家地里再说。

回顾一下个人数据存储的历史，可以发现网盘出现的必然性。20 世纪 70 年代随着个人计算机的出现，自然而然地产生了个人数据存储需要。只是当时计算机可以处理的内容非常少,一张 3.5 英寸的磁盘大约可以保存 1.4MB 的数据，人们使用计算机总是带着许多这样的磁盘。U 盘的出现，彻底终结了磁盘的历史。U 盘更加小巧，可以做成各种形状，读取的速度远远超过磁盘。现如今 U 盘也即将完成使命，将被网络硬盘所终结。

各类智能终端的出现，大大增加了人们在不同设备上交换数据的需求，如图 3-11 所示。U 盘在个人计算机之间传递数据还算方便。但是智能终端，尤其是手机，没有 USB 接口，无法使用 U 盘；苹果的 iPad 平板电脑，也不支持 U 盘。虽然如此，人们在智能终端设备上查看文件的需求却有增无减，迫切需要跨终端的文件共享方式，网络硬盘由此应运而生。美国风头最劲的网络硬盘公司名叫 DropBox，可以把这个名字演绎成“丢掉你的 U 盘”。

网络存储，将促进个人计算机和智能终端向两个不同的方向进化。个人计算机以强大的数据处理功能和运算能力，扮演信息、文件加工的角色，而形形色色的智能终端将适用于不同场景的信息展示和消费。以笔者的工作为例，仅在撰写研究报告或者书稿的时候，才会需要传统的笔记本电脑，去基金公司路演推介的时候，只需要携带一部 iPad 就够了，给客户展示的文件，早已事先保存在网络硬盘中了。

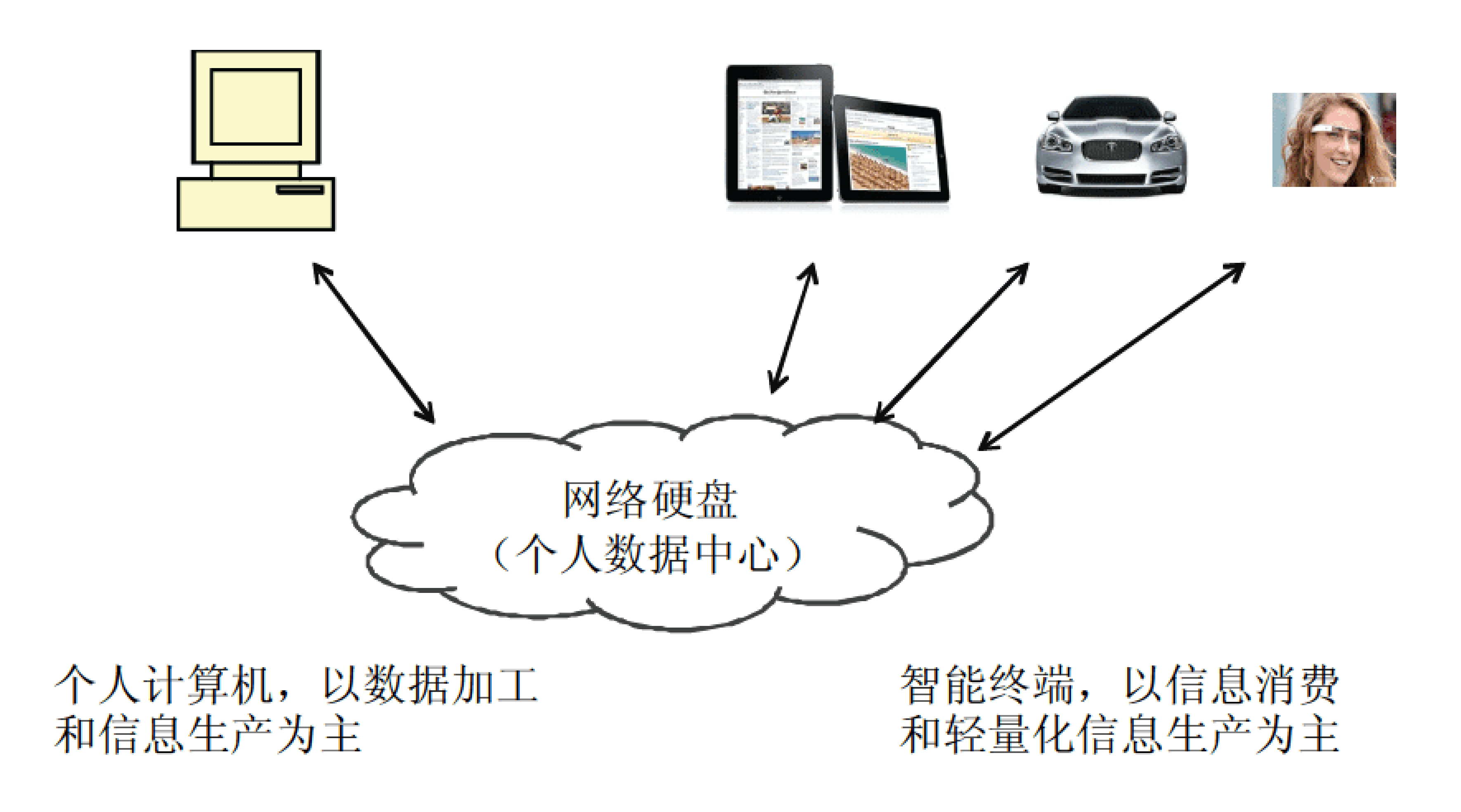


图 3-11 设备无关性的数据，引发产生重心的迁移

另一方面，智能终端上集成了丰富的传感设备，如摄像头、录音器、运动感应器、方向感应器等，可以随时随地采集大量的数据，包括照片、录音、录像等。这些数据，也需要方便、易用的保存和分享。网络存储的出现，提供了完美的解决方案。

因此网络存储一定是各大互联网公司、终端制造商、独立开发商的必争之地。联想公司通过网盘要打通联想旗下电脑、平板、手机之间数据交换和分享的通道，提高不同设备之间的协同性。短期来看，产生交叉销售的效应；长期来看，如果积累了丰富的数据，可以开展其他衍生业务。

中小企业市场，也是网络硬盘主攻的市场之一。海外 DropBox、谷歌等等早已开展了企业服务业务，国内如联想等也在向这个领域进军。这个市场刚刚兴起，对产业格局的影响，尚待观察。部分网络硬盘服务商及融资情况见表 3-3。

表 3-3 部分网络硬盘服务商及融资情况

名　　称	背景	用户规模	收费政策	投融资情况、估值
ASUS WebStorage	2008 年	400 万	免费+收费	
eSnips	——	——	免费+收费	200 万美元
Jungle Disk	——	——	免费+收费	2009 年 1 月被 Rackspace 收购
Windows Live Mesh	2008 年 4 月	300 万	免费+收费	微软

续表

名　　称	背景	用户规模	收费政策	投融资情况、估值
Mozy	2008 年 11 月	——	免费+收费	7600 万美元
DropBox	2008 年 9 月	4500 万	免费+收费	2.5 亿美元
Syncplicity	2006 年	很少	免费+收费	235 万美元
Wuala	2008 年 8 月	6000 万	免费+收费	被 LaCie 收购
ZumoDrive	2007 年		免费+收费	被摩托罗拉收购
Ubuntu One	2009 年 5 月	100 万	免费+收费	
iCloud	2008 年 7 月	200 万	免费+收费	2012 年 6 月 30 日停止 MobileMe 服务，原功能将整合至免费的 iCloud
box.net	2005 年	500 万	免费+收费	4800 万美元
115 网盘	2009 年	2500 万	免费+收费	约为 2000 万美元
酷盘	2010 年 7 月	800 万	免费+收费	2000 万美元
华为网盘	2009 年 4 月	2000 万	免费+收费	
金山快盘	2010 年 3 月	1000 万	免费+收费	融资计划考虑中
新浪微盘	2010年10月	1000 万	免费+收费	

案例：DropBox 简介[①]

DropBox 公司的创立人，2005 年刚刚从麻省理工学院计算机系毕业，名叫德鲁•休斯顿（Drew Houston）。2008 年 9 月，DropBox 的第一个版本上线运营，提供在线存储的服务。目前，DropBox 注册用户刚刚突破了 1 亿大关（见图 3-12），每天存储的文件超过 10 亿份。对此，休斯顿表示，自己感觉被用户赋予了极大的职责来帮助他们保存人生中最宝贵的记忆。而随着科技的不断发展，人们以后将可以通过各种设备访问自己保存在 DropBox 云端的数据。

腾讯科技援引国外传媒的一份报道，可以很好地说明 DropBox 带给人们的价值。以下段落引自腾讯科技，部分段落次序做了调整。

事实上，在 2011 年 4 月，DropBox 的注册用户为 2500 万，每天的文件保存量为 2 亿份；到 2012 年 5 月的时候，DropBox 也仅仅拥有 5000 万注册用户，每

① DropBox 的介绍，引自腾讯科技。

天的文件保存量为 2.5 亿份。但现在，DropBox 已经被安装到全球 200 多个国家和地区、支持 8 种语言的 2.5 亿台不同设备上。

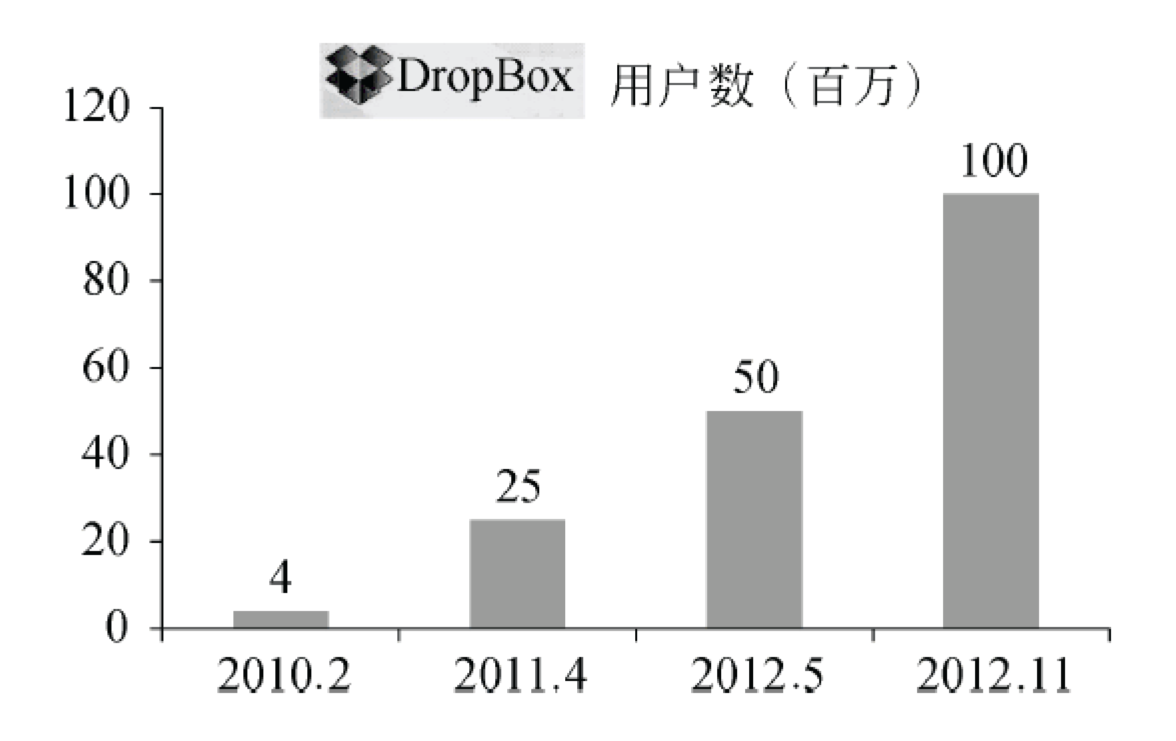


图 3-12　DropBox 用户数飞速增长

“1 亿的注册用户数是一个标志，这将 DropBox 放到了一个全新的高度，之前有少数的精英企业也曾经达到过这一水平。按照这一规模，当你能够帮助用户节省 10 分钟或者 1 小时，你就解决了人们一生中最大的难题，我们在这一领域才刚刚起步而已。”

作用巨大

“这给了我们一种感觉，感觉我们是在为整个世界解决难题，而不仅仅是硅谷地区。我们的用户多种多样，包括艺人、高中足球教练，甚至还有部分理论物理学家。但他们都表示，DropBox 的协作功能能够方便他们在全球分享实验数据，并帮助他们更好地开展研究。”休斯顿补充道。

除此之外，休斯顿相信，DropBox 正在帮助用户实现云计算时代到来时所做出的承诺，而 DropBox 则有望成为用户未来日子的开始。今后，如果笔记本电脑掉到水里，直接去苹果零售店再买一台就行了，因为 DropBox 可以帮助用户恢复所有数据，就像什么事情都没有发生过一样。

随后，休斯顿向人们讲述了一个令他颇有感触的故事：一位父亲用手机记录了

女儿出后五年的生活瞬间。但有一天，当他从洗好衣服的洗衣机往外拿衣服时，发现他竟然没有将手机从衣服中拿出来，手机中的女儿照片自然也全部无法读取了。幸运的是，这位父亲随后想起自己曾在手机上打开了 DropBox Camera 功能，因此他女儿的所有记忆都被完好无损地保存了下来。

发展迅速

DropBox 如今的规模和强烈的使命感已经吸引了大量用户。休斯顿透露，公司 2012 年年初的时候还仅有 90 名员工，但现在的员工数量已经超过了 250 人，这是 DropBox 员工增速最快的一年。休斯顿自豪地说："无论是什么职位，我们都能招到全球最顶尖的人才。想要一个工程师？没问题，iPhone 是谁设计的来着？"

举例来说，曾经的 Facebook 部门总监阿迪特亚 • 阿加瓦尔(Aditya Agarwal) 如今已成为 DropBox 的工程副总裁。阿加瓦尔曾帮助 Facebook 开发了"新鲜事 (News Feed)"和搜索功能，他是公司的第 9 号员工。阿加瓦尔表示："DropBox 是唯一一个让我能够再次产生这种影响的地方。"需要指出的是，能够吸引阿加瓦尔这样的优秀人才加盟公司只是 DropBox 成功融资 2.57 亿美元所带来的好处之一。

DropBox 有着非常巨大的宏伟蓝图，他们需要更多的人才来帮助自己达成梦想。尽管从笔记本电脑、手机和平板电脑，上传、下载、同步各种数据的任务看似简单，但其背后却需要很多复杂的工作。休斯顿表示："DropBox 有机会让你的智能手机、电视和汽车变得更智能。在这个领域，DropBox 将成为可以将一切连接在一起的一块帆布。"

未来蓝图

由于公司仍然保持独立运营的状态，因此 DropBox 有极好的机会成为一个数据层，就像 Facebook 现在所主导的社交层面一样。虽然苹果、谷歌和微软这些科技大佬都拥有自己的云存储系统，但它们却未必能够兼容其他厂商的设备。

“没有一家公司可以做到一切，但如果你拥有了 DropBox，你就可以不必担心设备的背后印着的是哪家的 Logo。”休斯顿如是说道。

DropBox 日后的覆盖范围完全可以延伸至冰箱、自动调温器和音响系统，只是目前这部分用户数量还相对较小而已。所以，为缺乏智能设备的新兴市场用户提供服务就成为了 DropBox 的一大使命。休斯顿认为：“目前全球有 20 亿网民，今后几年这一数字可能达到 50 亿。因此，任何拥有电脑或手机的人都需要 DropBox 这样的服务。”

不过，如果 DropBox 希望成为一个横跨多种设备，并可以为人们保存个人记忆和专业材料无所不在的数据层，DropBox 还有很长的路要走。

“在我们这个时代，无论你使用的是苹果的 Mac 还是微软的 Windows 系统都没有关系，因为我们会在另一领域继续创造奇迹，这是我们通往 10 亿用户大关的第一步。”休斯顿最后说道。

导读：

1. 多国报业公司破产、停刊传递了哪些信号？谷歌是互联网公司，还是手机制造商，亦或是电信服务商？千面谷歌的背后，其盈利的主要来源是大数据驱动的精准广告。数据量越大，谷歌的广告越精准，收入就越高。所以谷歌开创了一个精美绝伦的商业模式，免费地给大家使用各种优秀的服务，而谷歌只是拿走了大家使用的数据去找广告主收费。谷歌成功地推动大数据飞轮高速旋转，重压之下，以广告为生的报纸、杂志举步维艰也就不足为奇了。
2. 没有人比美国总统更善于营销，新的趋势是数据驱动的营销，它对奥巴马——美国历史上的第 44 位总统的续任起到了巨大作用，也是研究美国 2012 年大选中的一个关键元素。这同时也是一个信号，表明华盛顿那些基于直觉与经验决策的竞选人士的优势在急剧下降，取而代之的是数量分析专家与电脑程序员的工作，他们可以在大数据中获取信息，洞察选举形势。在政治营销领域，大数据的时代已经到来。
3. 传媒业市场规模庞大，且在高速成长。新技术、新商业模式、更多维度、更大规模的数据资产综合运用，将会催生下一个巨人。

第四章

大数据颠覆媒体行业

5000 亿美元的市场空间，将孕育下一家伟大的公司。

——笔者

“中新社柏林 12 月 2 日电，德国平面媒体业目前正经历着联邦德国建国以来最大倒闭潮。三家有影响的报纸月内连续宣告破产，造成上千人失业。两周前宣告破产的《法兰克福论坛报》带来的震惊尚未消除，《德国金融时报》也在日前宣告即将停刊，加上之前已消失的《纽伦堡晚报》，德国报业出现了有史以来规模最大的破产现象。”

与这些倒闭的报纸形成鲜明对照的是，谷歌一天赚一亿美元广告费的故事令人津津乐道。百度同样也有不俗的表现，2012 年，百度的广告收入超过了中央电视台。不久的将来，恐怕“中国第一媒体的王冠”不再仅仅属于 CCTV 了。

大数据时代，人们获取信息和传播信息的渠道、方法，都发生了根本性的变化。媒体行业正处在被颠覆的阵痛中。未来将是那些拥有大量数据资产的公司，掌握媒体的话语权！

如果把一家公司比成活生生的人，倾听和观察是人们成长的重要因素之一。对公司也同样如此，公司必须倾听消费者的心声，观察消费者的喜好，才能采取针对性行动。在中国文化的语境中，所谓察言观色，并不是一个褒义词，但是在商业战场上，这却是生存的不二法门。

如图 4-1 所示，宏观层面来看，企业有三个大的流程：第一，获取客户的信息，做到更充分地了解客户，也就是信息聚合；第二，企业内部流程，本章内容忽略不同企业内部的差异性，认为企业内部流程总是在外部的客户（消费者）的需求驱动的；第三，把产品或者服务传递给消费者（客户）的流程，也就是大家常说的营销。事实上，这三个流程是相互影响的，现实中很难完全分开。企业内部做设计的时候，可能随时去启动一些客户调查的工作；营销进行的过程中也可能会促进设计的变更，甚至生产工艺的改变；营销过程也往往离不开客户调查的工作。本章重点探讨大数据在信息聚合和精准营销过程中的价值，第八章将阐述大数据对企业内部组织、流程的影响。

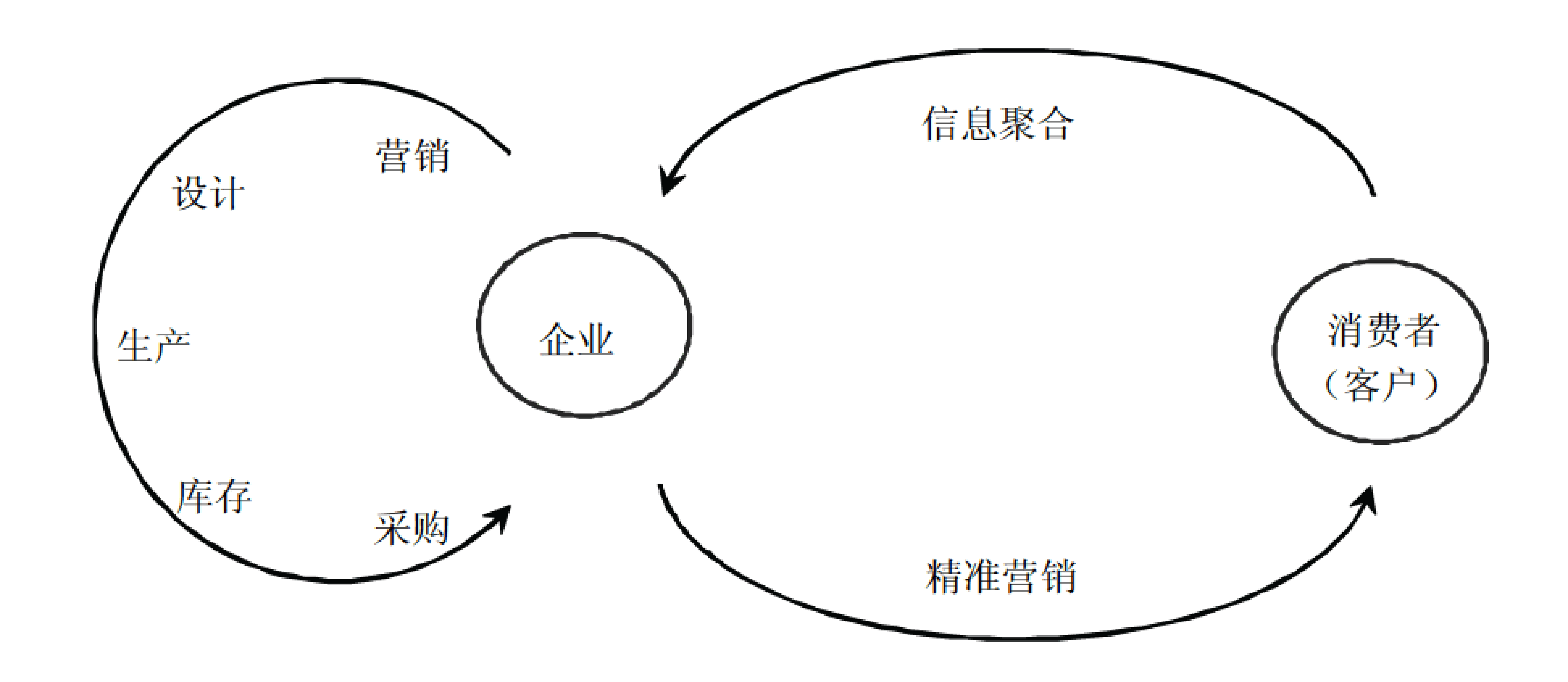


图 4-1　大数据令企业获取信息与传播信息的方式、方法发生了根本的变化

第一节　信息获取方式的变革——信息聚合

提要：

1. 偏狭的数据样本容易产生误导的结论。而利用大数据可以获得前所未有的精准预测能力。成功预测 2012 年奥巴马在全部 50 个州的选举结果的“书呆子”，其正是依赖广泛的数据来源。
2. 尼尔森联合 Twitter 推出新型的收视率调查工具。据此，可以清晰地看到市场调查行业发生的深刻变化。它们的数据来源更加丰富，不再局限于被动的监测，更需要积极主动地收集人们参与活动的数据。社交网络成为获取人们喜好的重要渠道。
3. 广开言路并充分利用社交网络中鲜活的数据，是提升政府治理水平的驱动因素。这个领域可以形成巨大的“信息聚合”产业，服务于政府、大型企业。

2012 年的美国总统大选，有一个“书呆子”上了各大媒体。34 岁的内特希尔沃凭借自己的“数学模型”准确地预测了这次大选全部 50 个州的选举结果。而在大

选日当天，他预测奥巴马将有 90.9%的可能性获得大半选举人票。他几乎打败了所有时政记者、政党媒体顾问和政治评论员。媒体称他为超级极客，“算法之神”，并认为其成功让所有“书呆子”扬眉吐气。

贝叶斯算法是统计学中的一个经典算法，内特希尔沃的数学模型也并无独创，但是这位“书呆子”建模分析的关键在于衡量某类数据的重要性。譬如，这些数据在历史上有何作用？又有怎样的偏向性？有没有其他的数据可以取代？

大数据时代，可供选择的数据来源极为丰富，人们也不再需要通过“采样”的方式来评估结果，而是有可能使用全部的数据进行运算，这无疑大大提高了预测的准确性。

偏狭的样本产生误导的结论

百事可乐曾经做过一个现场活动，让所有参加活动的人选择更喜欢百事可乐还是更喜欢可口可乐，结果是 90%以上的人都“更喜欢百事可乐”。为什么会出现几乎一边倒的调查结果呢？因为那些不喜欢百事可乐的人，根本就不会参加其组织的活动。所以，百事在“样本”选择上搞了一个“小花招”，从而得出“客观的”、“量化的”结论。

调查“样本”的选择，对结果的影响是至关重要的。在百事的样本中，无论采用多么先进的算法，都会得到百事胜出的结论。这也是为什么“几乎所有的时政记者、政党媒体顾问和政治评论员”都败在了“书呆子”手下的秘密。“书呆子”不是算法之神，那些记者们的数据来源，太过狭隘、太过主观，他们只希望自己期望的结果出现，就像百事操弄的活动一样。

样本量和算法的关系如图 4-2 所示。小的样本量非常容易被人为操纵，形成人们“希望”的结果。图 4-2 中第四象限就是典型的包装手法，拿一些不具有代表意义的数据，用“神秘”的算法包装，形成“导向”性的结果。如果有足够的数据，不需要什么特别的算法，就可以得到客观的结论。当然，如果辅以优秀的算法，人

们就可以拥有准确预测的能力，正如在本书概述中谈论的一样。[①]

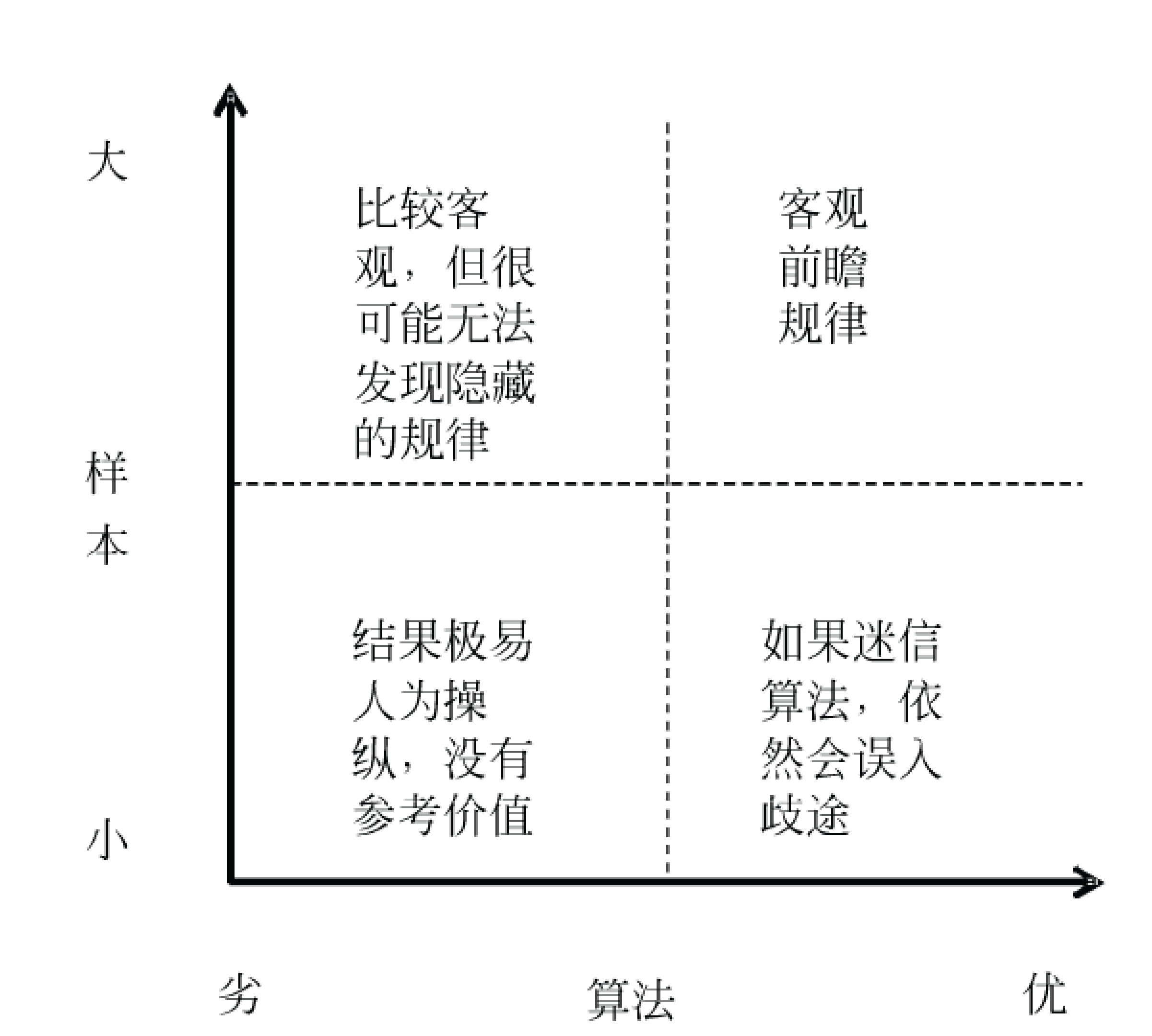


图 4-2　样本量与算法对结果的影响

尼尔森融合社交网络数据，推出新型收视率

传统上调查电视的收视率，需要通过电话访谈几百户人家。尼尔森市场调查公司是这方面的行家里手，据其官方网站介绍，尼尔森未来将获取更多的调查样本，对数字电视采用数据回传技术，可以在一个有线电视网内进行百万级别的普查。但是这远远还不够，最新的一则媒体报道，曝光了尼尔森和 Twitter[②]（推特）公司合作的消息。

新浪科技讯，北京时间 2012 年 12 月 18 日早间消息，美国市场研究机构尼尔森周一宣布，将携手 Twitter 推出新的收视率调查服务，监测 Twitter 上面有关某些电视节目的聊天内容。

这项新服务称为“尼尔森-Twitter 收视率”，将在 2013 年秋季推出，寻求监测

① 第一章　大数据预测未来的能力。
② Twitter 是一家美国公司，类似中国的微博服务提供商。

Twitter 用户在电视上观看 ABC 电视台“周一橄榄球之夜”、最新一季《国土安全》等节目的同时，在智能手机和平板电脑等“第二屏幕”上留下的评论和闲聊信息。

引用 NBC 热播节目《The Voice》执行制片人马克 • 博奈特(Mark Burnett)的评价：广告商应该重视那些可以激发观众在社交媒体上大量互动的节目。他表示，《The Voice》[①]之所以能在周二晚 18 岁至 49 岁的观众收视率中位列榜首，深度嵌入的社交媒体因素（如实时 Twitter 调查）起着至关重要的作用。博奈特说：“如果你是广告商，难道不想知道人们是在被动观看这台节目，还是积极参与这种观看体验？我认为 5 年以后，这种手段将让传统电视收视率调查变得过时。”Twitter 旗下媒体部门曾启动了一个为期一年的项目，旨在将“第二屏幕”的使用推向主流，而与尼尔森这样的知名市场调查机构的合作，无疑会推动 Twitter 在这方面的进程。

从尼尔森-Twitter 收视率，可以清晰地看到市场调查行业发生的深刻变化。它们的数据来源更加丰富，不再局限于被动的监测，更需要积极主动地收集人们参与活动的数据。社交网络成为获取人们喜好的重要渠道。如图 4-3 所示，尼尔森除了想方设法增加样本量外，也在大张旗鼓地拓宽数据来源的渠道，增加数据的维度。

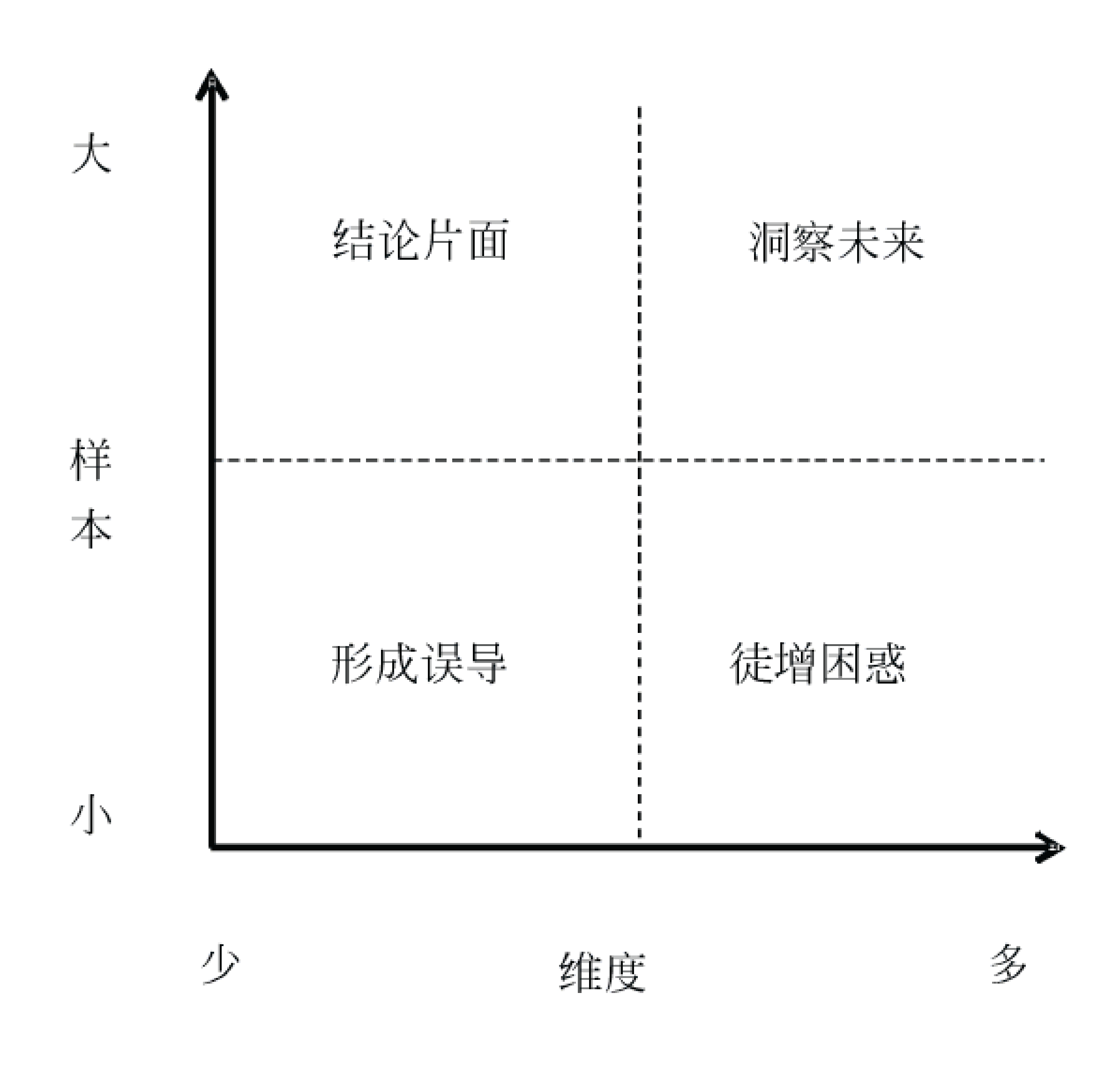


图 4-3　样本量与数据维度对数据分析结果的影响

① 类似《中国好声音》的一档节目。

社交网络数据，也是提升政府治理水平的重要载体

不仅是尼尔森等市场调查公司对社交网络中人们的言行感兴趣，政府机构也是另外一个大主顾。从古至今，公共舆论一直都是社会治理的一部分，是政府了解施政效果的重要渠道，也是人心向背的反映。因此，各级政府都会注重公共舆论。“舆情”是分析公共舆论中的话题热点、走势并及时提出应对策略的一种新型服务。

微博把人们聊闲天、传闲话、侃大山的天性发挥到极致。过往我们只能在街头巷尾，跟邻居们聊两句；现在，我们打个喷嚏，几亿人都能立刻知道。姚晨这位“微博女王”有近 3000 万粉丝，一句“早上好”，引发 2066 次转发，3739 条评论。这只是现在查询的数据，等到本书出版，这些评论和转发数肯定还会增加。

微博有巨大的传播和扩散效应，已经成为最重要的舆论场。中国人民大学舆论研究所经过研究发现，微博是 2011 年舆情事件的第一大信息源，占比达 20%以上。因此，微博也已成为各级政府密切关注舆论走向的“主阵地”。

舆情服务应运而生。目前，提供舆情服务的公司有很多家，上市公司中有拓尔思、人民网等。一般说来，如果省政府买了舆情服务，那么市政府一定会买。因为市里的事情，市长不想让省长先知道。以此类推，县政府甚至大一点的镇，都会跟风来采购舆情服务。现在一般的公司提供的舆情服务也比较简单，就是定期提供网上的热点分析，形成一份 Word 文档，就能坐地收银。

公司仅能提供基本的报告，将会在舆情服务的产业升级中落败。企业和政府在面临严峻的舆情形势时，需要的是舆情监测、舆情预警、舆情分析报告、应对处置、顾问咨询、舆情培训等一条龙式的服务。舆情服务的高级形式，应防患于未然，是从大众言论的蛛丝马迹中，发现可能形成的舆论热点，从而提前介入，利用恰当的方式方法引导舆论走势，而不能满足于做“事后诸葛亮”。再智慧的舆情应对处置都不如这个舆情事件不发生，所谓上医治未病就是这个道理。

如果想升级到舆情服务的高级阶段，没有大数据相关的技术，只能望洋兴叹。这一点也是中小舆情服务公司面临的难以逾越的技术门槛。舆情报告要求必须做到即时、全面且具备前瞻性。如果没有自建数据中心，没有大数据的采集、分析技术，没有成熟专业的舆情分析师团队，是不可能具备这种高质量的快速反应能力的。

如果从“信息聚合”的角度，把舆情当作一个产业来看待，其无疑蕴含着巨大的空间。消费者舆论中包括了大量的对产品、对公司、对品牌的意见和反馈，即便是不利于公司的负面舆论也是公司改变公众形象的契机。因此，无论是正面舆论还是负面舆论，都可能是公司潜在的客户，也都可能是潜在的广告受众。所以，从这个角度来看，类似像尼尔森这样的市场调查公司、拓尔思这样的舆情服务公司，都会获得前所未有的发展空间。事实上，舆情服务不仅仅是技术活，而是一个跨多个学科的综合工种。它不仅仅需要大数据，还需要社会学家、心理学家、传播学家、数据科学家共同努力，才能在这个领域有所作为。

信息聚合产业概览

根据用户研究的方法理论，用户数据信息获取来源可分为从用户的态度或用户实际行为中获取，研究的方法可分为按定性的直接方式获取或者按定量的间接方式获取。

根据具体项目的不同目标以及收集用户数据的难度，在用户数据获取的时候会选择不同的数据收集、聚合的研究方法。数据收集过程中剥离环境的方式如焦点小组、电话访谈等，基于实验室的数据获取方式如小范围眼动跟踪等。

现在在线媒体对用户数据信息的获取越来越强调对自然使用产品过程中机器所生成的海量结构化和非结构化数据的挖掘和分析，从用户主动填写或反馈的用户显性数据[①]和用户行为历史日志所反映的隐性数据[②]中提炼出有用的用户信息并进行信

① 显性数据指用户主动提供或反馈所形成的数据。
② 隐性数据是指不是用户主动提供，但是用户实际操作和任务执行过程中所形成的数据。

息聚合，如日志/摄像研究、留言板挖掘、数据挖掘分析、在线 A/B 实验等。通过一系列的数据预处理、分析和挖掘过程，建立模型，从而提取出有商业价值的用户信息，进而优化用户产品或者提高盈利能力。数字媒体盈利能力的核心竞争力主要体现在以下两个方面：第一是数字媒体本身所能收集聚合到的用户的各种显性和隐性数据的数量大小，这个涉及到数字媒体本身所覆盖的用户规模、用户粘性以及数字媒体的特点所生成的用户数据的差异等等。比如，社交网站 Facebook 在获取用户人口统计学信息和社交关系数据上有其他数字媒体所不具有的优势，即 Facebook 有全球规模最大的用户社交关系数据；电子商务网站亚马逊或淘宝在用户进行网上购物浏览和交易的数据上有其他数字媒体所不具有的优势，即亚马逊或淘宝全球规模最大的用户网上购物和商品偏好数据；搜索引擎公司谷歌在获取用户即时意图数据上有其他数字媒体所不具有的优势，即谷歌有全球规模最大的用户即时意图数据。这些媒体本身不同特性所导致的用户数据的差异，使得数字媒体本身所拥有的用户数据商业价值是不同的，后期对数据的利用方式也是不同的。第二是数字媒体在预处理、分析和挖掘用户数据上自身的能力不同。在这里说到的数据预处理、分析和挖掘的过程中，有大量与数据相关的问题需要解决，如可能的用户数据稀疏、海量数据的读写和存储性能等等。从原始的用户数据到最后数据价值被利用，涉及到一系列的数学模型设计和实现，公司之间的核心竞争力即体现于此。

围绕着用户数据跟踪获取、聚合和利用，现在美国已经形成了一系列完整的用户数据生态体系。信息聚合的数据来源有线下数据来源，也有在线数据来源。由图 4-4 可以看到，用户数据信息的生态体系中有数据提供商、数据交易市场、数据分析及用户定向提供商、数据管理、广告投放和效果跟踪的几部分参与者，每一部分都有典型的公司参与其中。



图 4-4 信息聚合产业生态图[①]

第二节 信息推送方式的变革——在线广告

提要：

1. 雅虎虽然开了在线广告的先声，但是谷歌充分挖掘了在线广告的商业价值，创造出了无与伦比的商业模式。传统平面媒体将不断消亡，在线广告将填补平面媒体倒闭空出的市场空间。
2. 直到以谷歌为代表的搜索广告的诞生，才解决了广告营销界的“哥德巴赫猜想”，即“我知道我的广告费有一半浪费了，但问题在于，我不知道是哪一半被浪费了。”这时，在线广告才散发出与线下广告不一样的独特魅力。
3. 数据资产的质量，是制约在线广告精准性关键的要素。

① 来源：波士顿咨询(BCG)研究 2012 年 1 月报告 “The Evolution of Online-User Data”。

互联网在线广告是精准营销的重要载体。

美国市场的一些统计数据表明，在线广告业务未来两到三年间将维持每年两位数的增长，预计到 2016 年市场规模可达 620 亿美元[①]。在线广告将不断挤压平面媒体，已经是主流的趋势。

图 4-5 是美国在线广告规模的增长变化图，截至 2011 年，美国在线广告规模已超过 300 亿美元。

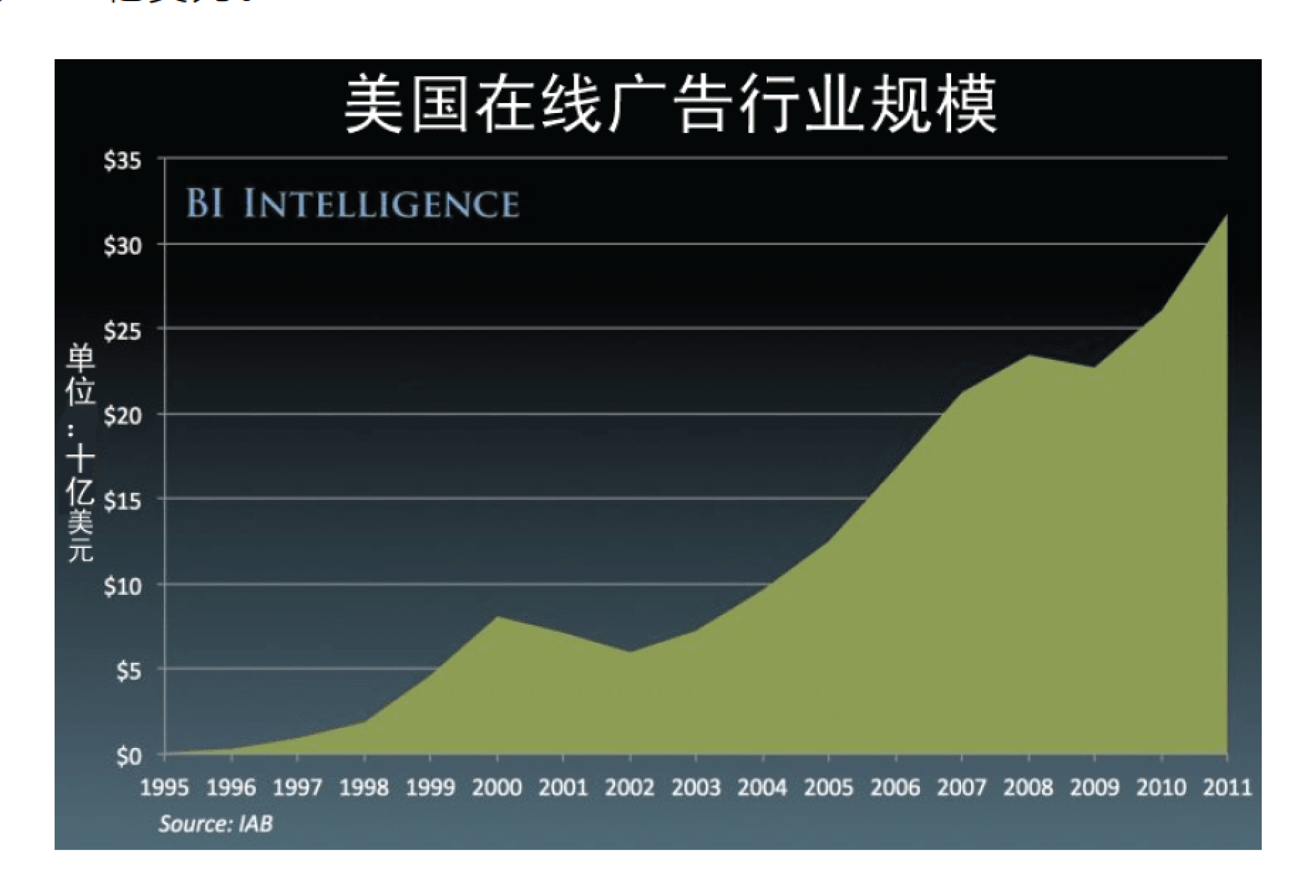


图 4-5 美国在线广告行业规模及增长趋势（数据来源：IAB）

图 4-6 是美国广告规模的增长变化图，截至 2011 年，美国在线广告规模已约占到美国广告总规模的 20%。

图 4-7 是美国主要媒体公司的广告收入增长变化图，可看到大的广告媒体公司的广告收入中，在线广告占比从 2006 年的 23%提升至 2011 年的 38%。

如图 4-8 所示，互联网上最早的广告仅仅是户外广告的翻版，典型代表是以雅

① 数据来源：eMarketer digital Intelligence，《Digital Ad Trends》，2012 年。

虎为首的门户网站。第二代是以谷歌为代表的搜索引擎，与之相对应的是互联网搜索广告，谷歌历年的收入中广告收入占 95%以上。第三代称为“内容广告”，根据网页内容的不同，而展示不同的广告。内容广告精彩纷呈，各路英雄大显身手，尚未形成绝对的垄断。第四代不妨称为“行为广告①”，这里并不是指艺术家们的“行为艺术”，而是消费者在互联网上的“行踪”。行为广告最能充分发挥大数据的预测能力，达到精准的广告效果。现在电子商务网站内部的“推荐系统 ”，就是行为广告的雏形。

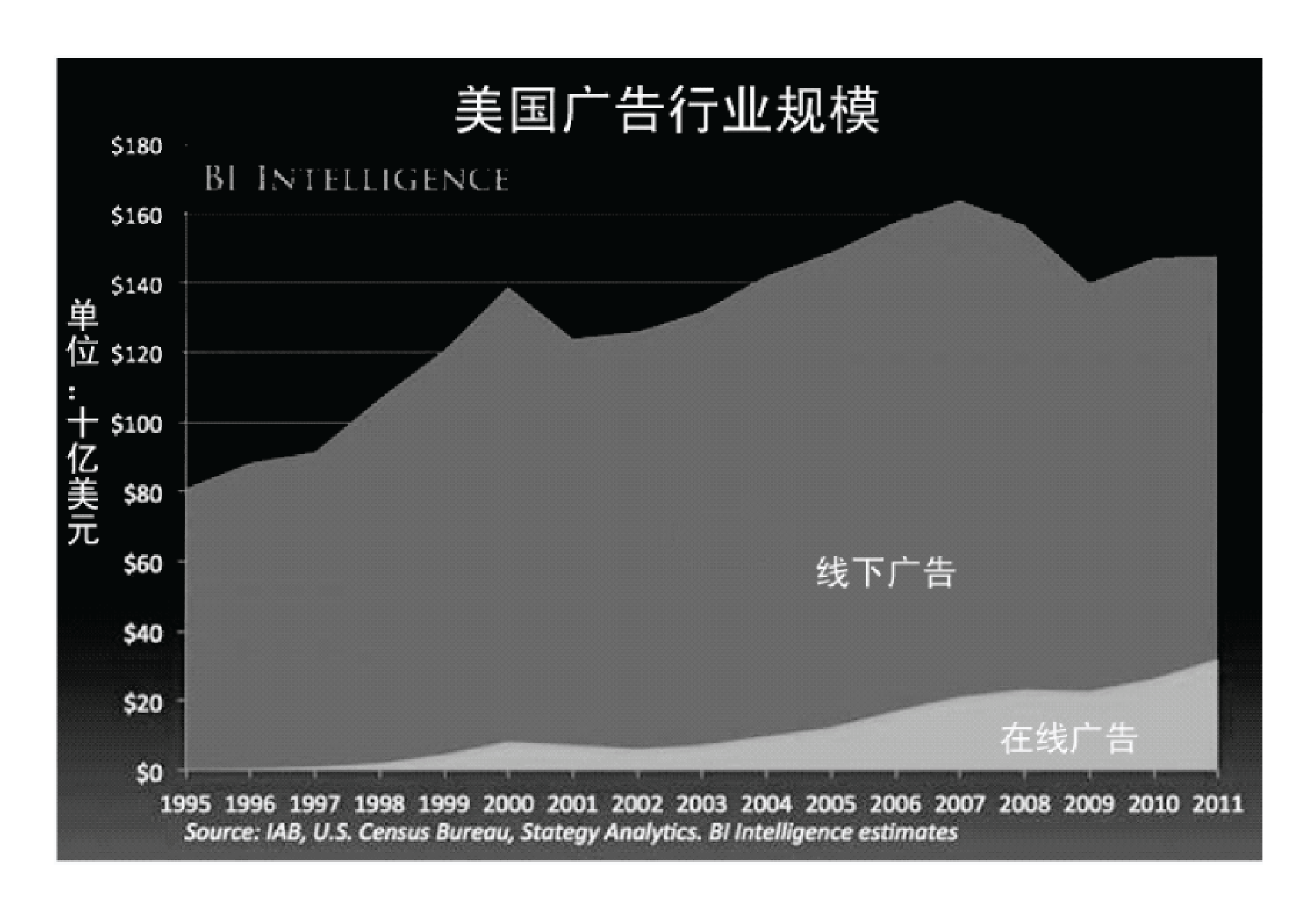


图 4-6　美国广告市场规模（数据来源：IAB）

在线广告市场发展非常迅速，以谷歌为例，在发展搜索广告、内容广告的同时，开发出了大数据处理技术，构成了现今火热 Hadoop 技术的基础。另外，广告理念、名词也在不断地推陈出新，从不同的角度来谈，可以有不同的广告类型。

① 搜索广告可以看成是行为广告的一个特例。它们的不同之处在于，当用户搜索时，已经有明确的、显性的需求；而用户在网络上的浏览、点击并不一定产生指向明确的需求，而是反映用户深层次、潜在的需求。在这个领域挖掘分析，可能会引领用户的消费。

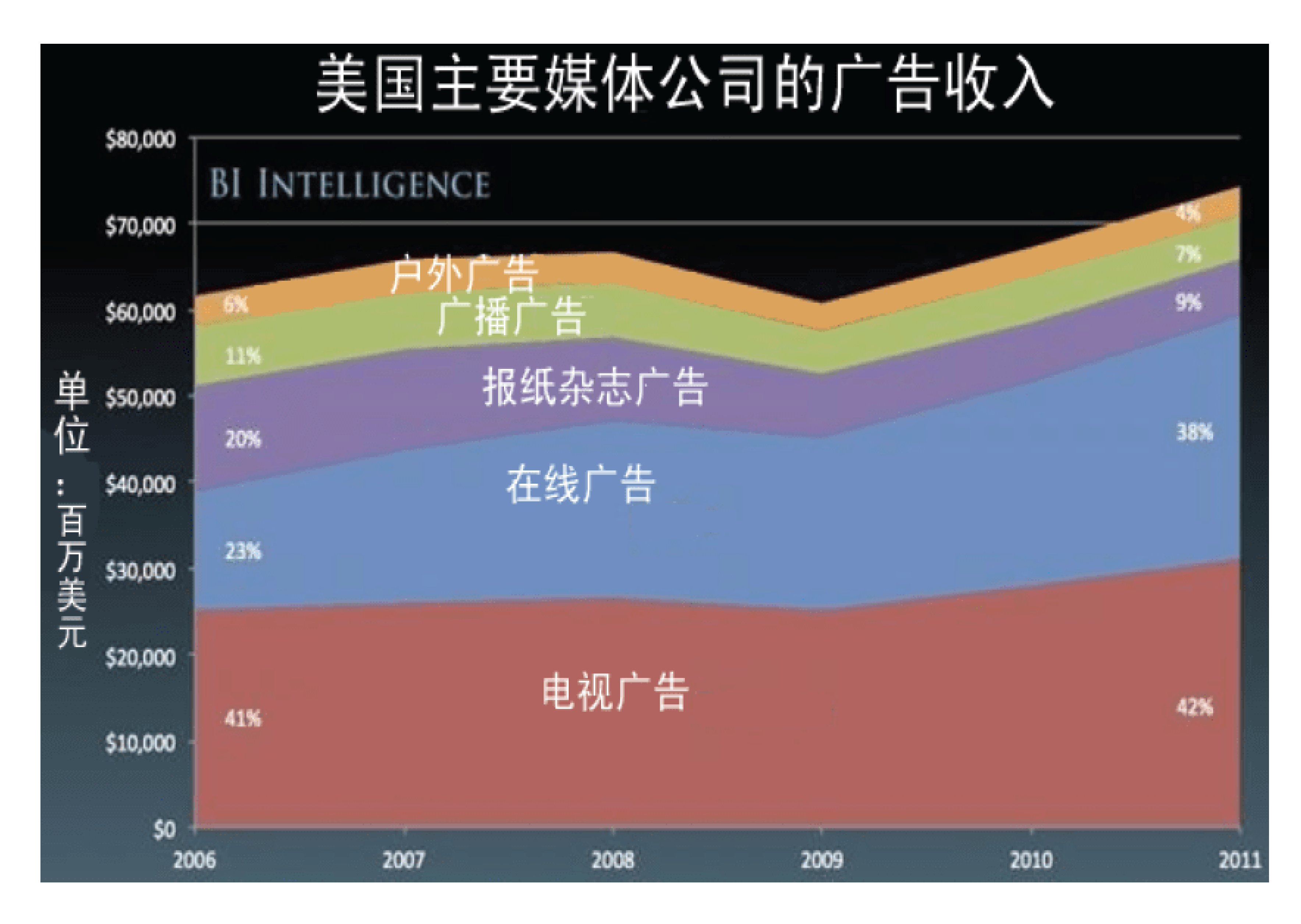


图 4-7　美国主要媒体公司的广告收入①（数据来源：IAB）

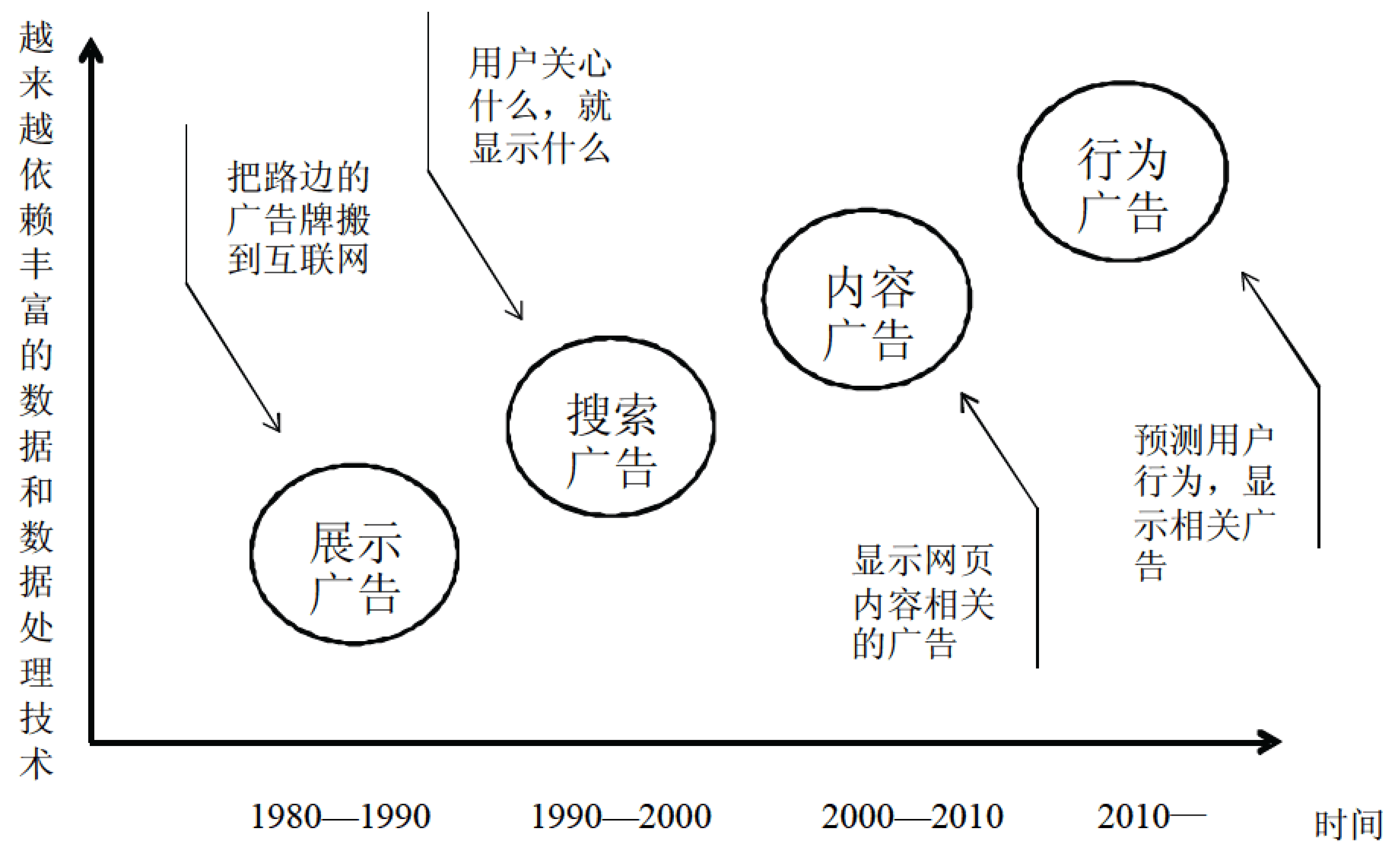


图 4-8　在线广告类别和演进，数据对广告的价值越来越重要

雅虎的贡献

《浪潮之巅》一书中用“英名不朽”来描写雅虎：“一百年后，如果人们只记得

① 主要媒体公司包括谷歌、雅虎、AOL、微软、Time Warner、Disney、Viacom、CBS、News Corp、NewYork Times 等。

两个对互联网贡献最大的人，那么这两个人很可能是杨致远和费罗（David Filo）。他们对世界的贡献远不止是创建了世界上最大的互联网门户网站雅虎公司，更重要的是制定下了互联网这个行业全世界至今遵守的游戏规则——开放、免费和盈利。正是因为他们的贡献，我们得以从互联网上免费得到各种信息，并且用它来传递信息，分享信息，我们的生活因此得以改变。”

杨致远在创立雅虎公司的时候，向大家描绘了一个非常诱人的商业图景。雅虎相当于互联网高速公路，依赖公路旁边伫立的广告牌来获利。如果大家都是通过雅虎上网，这条“高速公路”就成了机场高速，广告牌将大幅升值。雅虎创造性地提出了“门户网站”的概念，走上了一条提供优质内容、吸引访问流量、增加广告收入的良性循环之路。

伴随着雅虎公司等互联网门户兴起的即是互联网品牌广告的兴起。在线品牌展示广告随着 20 世纪 90 年代末互联网的发展和繁荣而兴起。很多年来，互联网广告创建、定价、包装和售卖的过程都维持不变，互联网品牌广告跟线下广告并没有任何本质上的区别，只是放置广告的媒体由线下媒体变为互联网媒体。

早期的互联网品牌广告并没有对流量进行切分以做到个性化精准投放，因此它没有解决约翰•纳梅克(John Wanamaker)曾提出的广告营销界的“哥德巴赫猜想”，即“我知道我的广告费有一半浪费了，但问题在于，我不知道是哪一半被浪费了。”直到以谷歌为代表的搜索广告的诞生，互联网广告才能计算出被浪费的那一半。这时，在线广告才散发出与线下广告不一样的独特魅力。

谷歌搜索广告——每天收入 1 亿美元

谷歌搜索引擎兴起、发展的历史，即是一部搜索广告的兴起及发展历史。谷歌现在已经成为全世界在线广告的霸主，拥有最完善的互联网广告产业链布局，也成为现在世界上市值最高的互联网公司。而谷歌历年的营收中，广告收入都超过了

95%的比例。

当大家在网上“冲浪[1]”时，身边的“广告牌”一闪而过，大家未必关心。这也是雅虎展示广告的最大困境，如图 4-9 所示。但是谷歌的搜索广告不同，大家每一次搜索，相当于告诉谷歌，“我正在找什么”，谷歌有可能根据大家当时的搜索关键字，提供相关的广告，如图 4-10 所示。

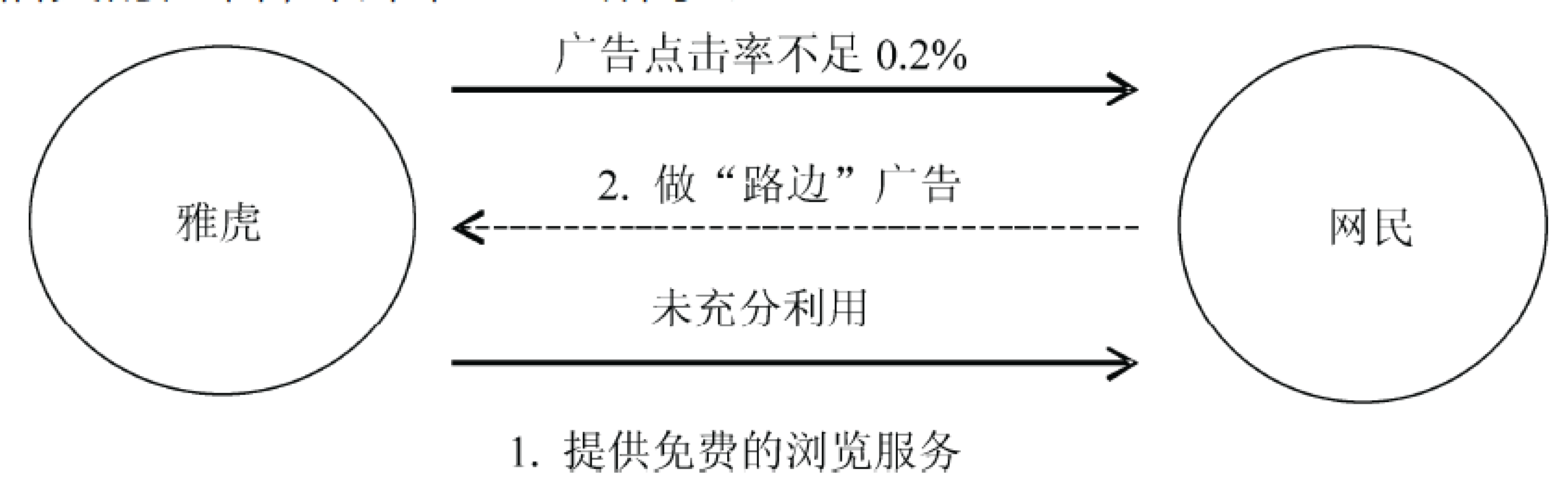


图 4-9　雅虎早期的展示广告商业模式

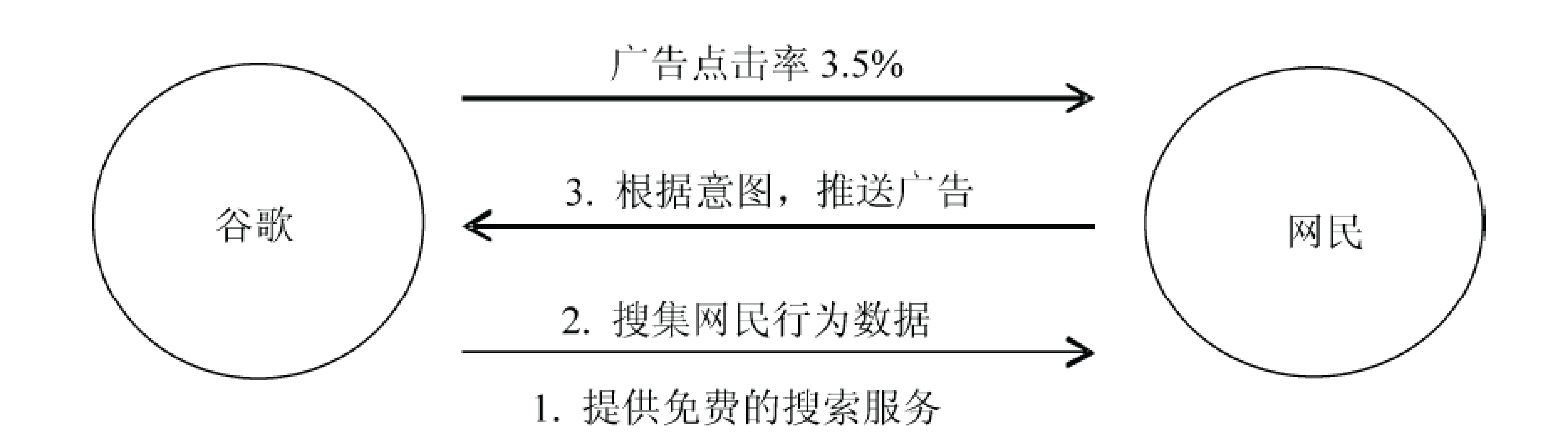


图 4-10　谷歌搜索广告的商业模式，搜索广告比现阶段展示广告的点击率高出一个数量级

譬如在谷歌的搜索框中输入“大数据”，搜索结果页面中，排在前三位的就是与“大数据”相关的广告，分别是“IBM 大数据处理解决方案”、“大数据解决方案——Intel”、“Splunk 大数据分析专家”。如何决定这三条广告的顺序呢？谷歌也有一套聪明的做法，哪个广告客户对“大数据”关键词的出价和广告质量分相乘的积最高，

① 在英文中，上网是“surfing the internet”，因 surfing 的意思是冲浪，即称为“网上冲浪”，这是一种形象的说法。

谁就排在第一位。相当于谷歌设计了一套“关键词”定价和广告质量评价机制，类似“拍卖”。百度开始是只对广告价格进行排序的，因此称之为“竞价排名”。因为仅仅考虑价格因素，一度导致虚假广告盛行，为业界所诟病。百度后期推出“凤巢广告”系统，也采取谷歌评估广告质量的机制，不再仅仅盯着“价格”这个单一因素。

根据谷歌公司 2012 年第三季度的经营数据显示，每天当网民使用谷歌的搜索引擎查询信息时，谷歌的搜索广告出现 56 亿次，网民每天点击广告的次数达到 1.9 亿次，每次点击给谷歌带来 0.53 美元的收入。尽管人们免费使用谷歌的搜索服务，事实上，在每次鼠标“咔哒、咔哒”声音中，钱就像流水一样跑到了谷歌的口袋！谷歌搜索广告占据美国市场绝对垄断地位见表 4-1。

表 4-1　谷歌搜索广告占据美国市场绝对垄断地位

2011	2012	2013	2014
谷歌	74.4%	77.9%	78.7%
微软	7.0%	7.0%	7.7%
雅虎	6.7%	4.5%	3.6%
美国在线	1.5%	1.0%	0.8%

注：数据来源 EMARKETER DIGITAL INTELLIGENCE 2012 “DIGITAL AD TRENDS”。

延伸阅读

世界上最大的拍卖，不是世界上哪一家老牌大规模的拍卖行所进行的行为，而是由成立至今才 15 年的谷歌广告系统所进行的拍卖。每时每秒，数以万亿的用户在使用互联网的搜索功能寻找有用的信息。而每一次搜索，带来的可能是不同广告客户对当次搜索行为所带来的搜索广告展示机会进行的拍卖竞价，这些拍卖行为完全由计算机完成，每时每秒都给谷歌带来巨大的收入。谷歌的搜索广告与以雅虎为代

表的展示广告有本质的不同，其最主要区别在于：谷歌的搜索广告是由计算机根据广告客户竞价、广告本身质量及与用户搜索查询的相关性等多因素来共同起作用，将满足相关性和盈利两方面总体价值最大化。也就是说，每一条谷歌广告都是由计算机算出来而放置在合适的广告位上的，而不是像早期雅虎的品牌广告一样，是人为放置在媒体上的，因此搜索广告也称为计算广告。

在线计算广告是技术和产品驱动的，是技术和计算的导向，而原有的在线品牌广告是创意和客户关系驱动的。计算广告出现的根本原因，在于数字媒体的特点使得在线广告的精准定向成为可能，即给不同用户展示不同广告。传统媒体上需要实现“用户定向的精准投放”的成本特别高，在线媒体上进行精准投放成本很低。在线广告竞价系统和实时竞价系统里，在线广告的计算已经成为核心问题。在线广告和传统线下广告的最大区别之一还在于在线广告的即时效果可以更好地衡量，因为广告是可以点击的，而点击后用户行为也可以较方便地跟踪获取，因此获取广告即时效果的数据更加容易，精准定向投放和即时效果方便衡量促进了在线广告的标准化和迅速发展。

广告的研究人员把用户、广告客户和媒体资源的交互过程来建立模型进行研究，以使这三者总体的收益最大化，其中两个比较重要的模型是广告的投放模型和拍卖模型。投放模型是解决广告客户投放广告给哪些目标用户的问题，比如一个卖运动鞋的广告客户会希望能选择广告受众用户，以便有选择地把他的广告定向投放给那些可能对他的鞋感兴趣的用户，这样可以节省广告投放费用而提高广告投放效率。其在搜索广告上的实现方式是，这个卖运动鞋的广告客户购买一些与运动鞋相关的关键词，通过与用户搜索查询相匹配的方式来找到他的目标用户。而拍卖模型是解决多个广告客户如何竞争针对同一目标用户的广告展示机会的问题，比如有一个对

运动鞋感兴趣的用户访问有广告的页面，同时有几十个广告客户希望把广告投放给这一个用户。但是因为广告位是有限稀缺的，如何拍卖有限的广告位，以达到用户、广告客户和媒体资源的收益最大化，这一问题涉及到非常多的用户和广告客户数据信息。计算广告系统在 100 毫秒的时间内，通过计算找出展示在合适广告位的合适广告。这一数据量非常巨大，只能是通过计算机计算完成。

同时，因为用计算机实现投放和竞价，搜索计算广告系统能容纳比互联网品牌广告系统更大规模的广告客户，搜索计算广告使得在互联网投放广告的客户规模远远超过在线品牌广告的客户规模。谷歌的广告客户有几十万家，而且广告客户越多，对计算广告的收益和精准性越好，这样使很多中小规模的客户也能在互联网投放广告，而这是以雅虎为代表的门户网站品牌广告所无法实现的。广告客户规模的增加，以及对海量用户和媒体资源的数据信息的利用，都使得大数据技术在计算广告系统中可以大展身手。

内容广告

微软公司在 2012 年 6 月 7 日，宣布在自己的浏览器 IE10 中，默认开启“Do Not Track”（不被追踪功能），引起美国广告业的轩然大波，以至 IE10 浏览器都没有获得万维网联盟[①]的承认。微软这个举动，仅实施了短短六天即惨遭夭折。

微软和广告业的争端，让一项浏览器追踪技术，大白于天下。事实上，人们用于上网冲浪的浏览器，通过 Cookies[②]忠实地记录下，大家曾经到访过哪个网站等信

① 万维网联盟（World Wide Web Consortium，W3C），又称 W3C 理事会，1994 年 10 月在麻省理工学院计算机科学实验室成立，建立者是万维网的发明者蒂姆·伯纳斯·李。万维网联盟是国际著名的标准化组织，自 1994 年成立后，至今已发布近百项相关万维网的标准，对万维网发展做出了杰出的贡献。

② Cookie 是计算机术语，中文名称为小型文本文件或小甜饼，指某些网站为了辨别用户身份而存储在用户本地终端（Client Side）上的数据（通常经过加密）。

息。而一些广告商则利用保存在 Cookies 中的用户访问记录，来判断人们的喜好，推送更加精准的广告。通常情况下，浏览器默认开启 Cookie 功能，也就是浏览器默认记录用户上网的行踪，加大了用户隐私暴露的风险。

IE10 的“Do Not Track”功能相当于告诉网站，这个用户不希望被追踪。如果网站明确获知用户的意愿是“不”，而依然追踪用户的行踪，就属于非法形为。因此 IEID 遭到强烈的反对和抵制。

这场争端以广告业的胜出告一段落，从中人们也约略了解了在线内容广告行业的尴尬处境。为了展示更加具有针对性的广告（譬如，如果了解用户刚刚从一个汽车网站跳转来看新闻，就可以针对性的提供汽车广告），而不得不更多了解用户的喜好。但是普通的网站除了 Cookies 似乎并无良策，谷歌的 AdSense 技术另辟蹊径，自成一家。

谷歌的 AdSense 技术

谷歌公司的 AdSense 产品为提升内容广告的点击率开拓了一个思路。AdSense 通过分析网页的内容，提供和内容相关的广告。隐含的假设是，广告和用户正在浏览的网页内容关联度越高，用户关注的可能性越高。

百度百科有一段形象的原理说明，摘录在此（略有改动）。

谷歌 AdSense 原理形象说明：

1. 在网页中加入一小段谷歌提供的 AdSense 代码；
2. 用户浏览该网页；
3. AdSense 代码对谷歌广告服务器说：“嘿,给我一些广告”。
4. 谷歌广告服务器回答说：“不行，谁知道你页面里有什么东西啊？”
5. 用户于是看到一个没有谷歌广告或者带着谷歌公益广告的页面；

6. 谷歌广告服务器派出一个机器人浏览这个网页；

7. 服务器分析网页的内容，发现“比萨饼”这个单词出现了 20 次，“华盛顿”出现了 6 次；

8. 于是服务器认为这个网页在讨论“华盛顿的比萨饼”。

9. 又有用户浏览该网页；

10. AdSense 代码对谷歌广告服务器说：“嘿，给我一些广告”。

11. 谷歌广告服务器回答说：“好，这是个关于华盛顿比萨饼的页面，给你一些华盛顿比萨饼外卖广告吧!”

12. 用户心想“嗯，正打算叫比萨外卖呢”，点击广告；

13. 这样你赚了一点点钱；

14. 从第 9 条开始周而复始。

AdSense 必须“理解”网页的内容，这就需要一些统计算法，如关键词处理、语义分析之类。这也是目前大数据应用的一个热点领域。

谷歌的邮件系统 Gmail 也利用了类似 AdSense 的技术。当使用网页版的 Gmail 服务时，邮件正文右侧会显示一些文字广告，不同的邮件内容会显示不同的广告。谷歌 AdSense 技术促使内容广告的点击率不断提升，但是相比搜索广告，还是有数量级的差距。后文会分析这种现象的成因，推测谷歌下一步的发展方向。

图片识别广告——Pixazza

Pixazza 是一家图片匹配广告服务商，被称为 AdSense for Images，只不过 AdSense 只能匹配文字，而 Pixazza 则可以匹配图片中引入注目的商品，就像大部分人在商场购物，都会被模特展示的衣物吸引一样。如图 4-11 所示，如果用户喜欢图中女模特的帽子，或者男模特的上衣，只需要将鼠标悬停对应的星标上，就

能看到更多信息和类似产品的价格，当然最重要的还有一键购买。

图 4-11 Pixazza 图片内容广告

搜索广告与内容广告的对比

谷歌公司占据美国广告市场的垄断地位，而且火热的大数据处理技术 Hadoop 也是来源于谷歌公司的工程实践。所以这里用谷歌的数据为例，来分析搜索广告和内容广告的差距，探究内容广告可能的发展方向，剖析大数据在内容广告市场的应用和前景。

我们不去赞叹谷歌“日进斗金”的商业模式，而把注意力放在搜索广告和内容广告两类对比鲜明的数字上。

每天谷歌在各类网站上投放的内容广告远远高于搜索广告，但是相比搜索广告带来的收入，内容广告几乎可以忽略不计，主要原因就是没有人愿意点击跟自己无关的广告。人们在利用搜索引擎查找资料的时候，相当于把自己当时的“愿望”告诉了谷歌，谷歌知道人们要干什么，及时“奉上”广告，这时广告的精准性远远高

于漫无目的的内容广告。尽管谷歌采用了 AdSense 等内容相关技术，但是依然在预判用户行为方面稍逊一筹。

简单解释一下广告点击率和转化率这两个概念，它们是在线广告行业的核心指标。所谓点击率，就是用户点击广告的百分比。广告内容越贴近用户需求，点击率就会越高。点击率的高低直接反映了公司对用户的理解程度，如果对用户了如指掌，理论上讲广告就会百发百中。理想很丰满，现实很骨感。谷歌搜索广告的点击率仅仅在 3%～4%之间波动，百度搜索广告的点击率也在同样的数量级。这个指标可以衡量公司技术实力。但是需要注意一个误区，谷歌点击率是在数十亿的样本中计算得到的，如果某家公司的广告点击率非常高，判断其是否靠谱的一个办法，就是检查它的样本量。如果样本量过小，这个指标就失去了应有的价值。

转化率[①]是指用户点击广告链接进入产品页面后，产生了购买行为的比例。这个指标更多衡量广告客户的实力，与广告公司的关系不大。所以，从表 4-2 的数据中，可以看到谷歌的内容广告和搜索广告的转化率差别并不明显。

表 4-2　谷歌搜索广告和内容广告数据对比表

	搜索广告	内容广告
广告每天显示数	56 亿	242 亿
广告点击率	3.47%	0.18%
平均每天点击数	1.93 亿	0.45 亿
广告转化率	5.63%	4.68%
平均单击价格	$0.53	$0.35
平均每天收入	1 亿美元	1600 万美元

注：数据来源 HTTP://WWW.WORDSTREAM.COM/BLOG/WS/2012/10/25/GOOGLE-FACES，依据谷歌公司 2012 年第三季度数据整理而成。

如果内容广告能够把点击率提升一个百分点，对谷歌而言就会增加近 9000 万美元的收入。毫无疑问，所有的广告公司都把提升内容广告的点击率作为主要的战

① 转化率的定义各个广告客户不统一，有的是以用户注册作为转化，有的是以用户下载作为转化等等，这里是指“购买转化率”。

略方向。点击率的提升与对用户的理解息息相关。要想更理解用户，离开大数据技术，则是不可能完成的任务。

谷歌的成长史，也是一部收购史，截止到 2012 年 11 月 30 日，谷歌已经收购了 121 家企业。谷歌的收购非常明确，要么获得新的提升广告精准性的技术，要么获得新的数据来源，增加数据资产的维度。这里仅通过几家典型的收购案，来观察谷歌是如何不断充实其“数据资产”的。回顾一下图 3-1 所示的数据资产评估模型，再看表 4-3。

表 4-3 谷歌公司的典型收购

收购日期	公司	性质	改造/整合对象	数据资产
2003 年 2 月	Pyra Labs	博客软件	Bloger	收集博客数据
2003 年 4 月	Applied Semantics	网络广告	AdSense、AdWords	
2004 年 7 月	Picasa	图像管理工具		收集图像数据
2004 年	ZipDash Where2 Keyhole	地图	谷歌地图	这三家公司丰富的地图数据、获得分析技术
2005 年 7 月	Current	宽带因特网连接	网络骨干	基础电信数据
2005 年 8 月	Android	移动设备操作系统		获得移动设备使用数据
2006 年 3 月	Upstartle	文字处理器	Google Docs	获得文档数据
2006 年 10 月	YouTube	视频分享网站		获得视频数据
2006 年 8 月	Neven Vision	人脸识别	Picasa	图像挖掘技术
2007 年 4 月	DoubleClick	网络广告	Adsense	广告技术
2007 年 6 月	Panoramio	照片分享		获取图像数据

续表

收购日期	公司	性质	改造/整合对象	数据资产
2010 年 5 月	Simplify Media	音乐同步	Android	获取音乐数据
2010 年 5 月	Ruba	旅行向导		获取出行数据
2010 年 8 月	Slide.com	社交游戏	Google+	获取娱乐数据
2010 年 7 月	ITA	航班信息	GoogleFlight	获取出行数据
2011 年 8 月	摩托罗拉	智能手机		控制产业链

注：数据来源维基百科。

谷歌的新动向——提供光纤接入互联网服务

谷歌的光纤接入服务速率达到 1000Mbit/s，相比国内电信运营商提供的缩水“宽带”，真是一个在天上，一个在地上。问题是，谷歌作为广告商，为什么会提供光纤接入服务，介入基础电信运营领域呢?

在本书第三章表 3-1 中，分析了电信运营商的数据资产特点，其中蕴含了谷歌提高内容广告点击率的重要资产。电信运营商拥有丰富大量的人们上网记录和通话记录，尤其在移动互联网时代，这些数据真实、直接、随时随地地反映了用户潜在的需求。如果结合大数据处理技术，谷歌很可能在内容广告领域，再现搜索广告领域的风光，创造一个新谷歌。

网友提供了利用电信信令数据精准定位人群的案例。这些条件完全取自用户手机的通话特征，甚至不需要通话，利用手机和基站间联络的信号，就可做出判断。例如:

体育场观众的判断条件：① 体育活动起止时间；② 体育场基站覆盖区域的用户；③ 在活动期间出现在该体育馆基站范围内停留时间大于 1 小时且小于 8 小时；④ 在活动开始前 3 小时到活动开始后 1 小时之内出现在体育场基站覆盖区域；⑤ 在

活动结束前 1 小时到结束后 2 小时内离开体育场基站覆盖区域；⑥ 选择在网时长超过 2 个月的用户；⑦ 不是警务通套餐用户……

机场离港客户的判断条件：① 选择在机场基站区域覆盖的客户；② 一个月内在机场区域内出现的累计时长小于 50 小时；③ 一个月内在机场区域内出现的累计天数小于 10 天；④ 3 个月月平均 ARPU 大于 50 元；⑤ 手机用户在机场区域内关机……

机场来港客户的判断条件：① 选择在机场基站区域覆盖的客户；② 一个月内在机场区域内出现的累计时长小于 50 小时；③ 一个月内在机场区域内出现的累计天数小于 10 天；④ 手机用户在机场区域内开机并离开机场区域，且停留时间小于 120 分钟；⑤ 3 个月平均 ARPU 大于 50 元……

精准定位不同的人群是提升内容广告点击率的法宝。谷歌正是瞄准了电信运营商这些宝贵的数据资产，才悍然介入基础电信运营领域。不排除谷歌未来收购电信运营商的可能性，这个领域同样存在本书第六章指出的行业垂直整合的趋势。运营商应向媒体转型，而媒体也会介入运营，这个趋势是不可扭转的。

第三节　行为广告领域将孕育“新谷歌”

提要：

1. 亚马逊的推荐系统是行为广告的典型代表。亚马逊的广告理念就是根据人们的不同喜好，推荐合适商品。无须促销，不同的人，登录亚马逊网站看到不一样的主页，不一样的商品。

2. 多维度“数据资产”的争夺是行为广告的主战场。谷歌公司发布的智能手机、优酷视频、在线文档都是增加数据资产维度的手段。Facebook 公司将谷歌拒之门外，主要原因就是 Facebook 的数据资产蕴含巨大的广告价值，马克·扎克伯格准备自己发掘这座金山。

3. “生态系统”的培育是问鼎行为广告的基石。大数据技术是决胜行为广告市场的屠龙刀。

行为广告是杜撰的名词，言下之意是充分利用人们在网上留下的各种行踪，精确预测个人需求，从而推送更加精准的广告。行为广告的呈现形式，还是以现在的内容广告为主，但将是高度个性化的“内容广告”。

亚马逊电子商务网站的推荐系统，为人称道不已。如果你是数码控，登录亚马逊网站首页看到的几乎都是数码产品；如果你是一位新妈妈，则网站首页将会是满满当当的婴幼儿产品。这是亚马逊公司根据网站上积累的大量的用户购买、浏览信息，做出的个性化调整。这类推荐系统，就是行为广告的雏形。谁能把这种思想率先推广到整个互联网，谁就将成为下一个谷歌。

可以把行为广告看成是内容广告的高级阶段。目前，内容广告市场相较于搜索广告市场而言，竞争格局尚存在变数，不存在绝对垄断的公司。美国最大的内容广告媒体谷歌联盟和 Facebook 展示的广告市场份额均未超过 20%，见表 4-4。

表 4-4　美国内容广告市场格局及发展预测

	2011	2012	2013	2014
Facebook	14.0%	16.8%	17.7%	17.1%
谷歌	13.8%	16.5%	19.8%	21.7%
雅虎	10.8%	9.1%	8.1%	7.5%
微软	4.5%	4.4%	4.3%	4.4%
美国在线	4.3%	4.0%	3.8%	3.7%

注：数据来源 EMARKETER DIGITAL INTELLIGENCE 2012 “DIGITAL AD TRENDS”。

亚马逊公司的推荐系统是行为广告的雏形

互联网曾经流传过这么一件真实的案例：在美国明尼阿波利斯市，一个中年男子怒气冲冲地走进塔吉特（Target）百货要求见经理，手里拿着这家商场寄给他女儿的优惠券。“我女儿收到了这些！”他对着商店经理咆哮，“她还在上高中，你们就寄给她母婴用品的广告？你们是在鼓励她怀孕吗？”

经理一头雾水，他看了看邮件，里面确实有寄给这个女孩的孕妇服和婴儿床的广告单。经理不得不反复向这位中年男子道歉，事情才得以平息。

过了几天，塔吉特百货的经理又打电话给这位父亲，想表示歉意。但在电话里，女孩的父亲说话吞吞吐吐，显得很尴尬：“我和女儿长谈了一次，我没有察觉到家里的一些事——她确实怀孕了，我向你们道歉。”

塔吉特百货公司是如何早于父亲知道一个女孩未婚先孕，从而向女孩推荐母婴用品优惠券的呢？购物商店如何比用户自己还更了解他们究竟需要什么，从而向用户推荐商品和优惠券广告的呢？说到这些，不得不提及商品推荐系统及推荐系统的鼻祖亚马逊。

亚马逊是世界上最大的网上商店，成立于 1995 年，是一家眼光长远的伟大公司。亚马逊最奇特的地方在于，自公司成立以来就开始亏损，而且一年比一年亏得多，仅 2000 年一年净亏损就达到了 14.1 亿美元，2000 年之后亏损的步伐才有所放缓。亚马逊从成立开始一直连续亏损了 8 年时间，终于在 2003 年第一次开始盈利，净利润第一次由负数变为正数。亚马逊的创始人贝索斯是一个眼光很长远的人，在亚马逊成立以来经历了互联网泡沫的形成、滋长和破灭，历经太多投资人和投资机构的做空，但是贝索斯总是我行我素。在亚马逊公司年报每年一度的致股东的信里，贝索斯总是把 1997 年他第一次给股东的信重新贴一遍，其中老是强调“It’s All About the Long Term”，即所有亚马逊所做的都是关乎长远的事情。

在 2010 年贝索斯给股东的信中，一开始便提到“如果你走进亚马逊的某些会

议，你可能会觉得你走进了一个计算机科学的讲座。在现在关于软件架构的教科书中，已经很少能找到亚马逊没有应用的模型。我们使用高性能交易系统、复杂的渲染和对象缓存、工作流和排队系统、商业智能和数据分析、机器学习和模式识别、神经网络和概率决策及很多其他技术。”在亚马逊的这些计算机技术中，非常重要的一部分是应用于亚马逊的推荐系统。

2012 年前三财季，亚马逊营收达到了 398.3 亿美元，与 2011 年同期的 306.5 亿美元相比大涨了 30%。亚马逊能有如此惊人的营收增长，其推荐系统功不可没。在亚马逊商品爆炸的网上商店，让用户发现自己潜在的需求，对于亚马逊是至关重要的，它已经将推荐的思想渗透在应用的各个角落，深度整合到购物流程的方方面面，从商品发掘到结账付款，几乎无处不在。登录亚马逊.com，会看到许多商品推荐板块，点击进入某个商品的网页，“人气组合”与“（浏览了该商品的）用户还购买了其他商品”等栏目赫然在目，这一切都使亚马逊的用户惊叹于为什么这个网上商店总是能猜到自己到底想要些什么。

亚马逊推荐的核心是通过数据挖掘算法和比较用户的消费偏好与其他用户进行对比，借以预测用户可能感兴趣的商品。亚马逊采用的是分区混合的机制，并将不同的推荐结果分不同的区显示给用户。

亚马逊利用可以记录的所有用户在站点上的行为，根据不同数据的特点对它们进行处理，并分成不同区为用户推送推荐。后文会把推荐方式进行整理再介绍典型的推荐方式，这里先介绍亚马逊典型地向用户推荐的表现形式。

今日推荐 (Today’s Recommendation For You)：通常是根据用户近期的历史购买或者查看记录，并结合时下流行的物品给出一个折中的推荐。

新产品的推荐（New For You）：采用了基于内容的推荐（Content-based Recommendation）机制，将一些新到物品推荐给用户。在方法选择上由于新物品没有大量的用户喜好信息，所以基于包含共同特征的物品和内容推荐能很好地解决这个问题。

基于用户浏览过的商品的推荐（(Recommended Based on Your Browsing History)：基于用户以前浏览过的商品的特征及相关的商品推荐给用户，这也是基于包含共同特征的物品和内容推荐方式。

其他人正在浏览的商品（What Other Customers Are Looking At Right Now）:通过具有类似用户特征的用户正在浏览的商品信息进行推荐，这是基于用户维度的推荐。

捆绑销售 (Frequently Bought Together)：采用数据挖掘技术对用户的购买行为进行分析，找到经常被一起或同一个人购买的物品集，进行捆绑销售，这是一种典型的基于物品或内容的协同过滤推荐方式。

别人购买 / 浏览的商品 (Customers Who Bought This Item Also Bought)：这也是一个典型的基于项目的协同过滤推荐的应用，通过社会化的机制，使用户能更快更方便地找到自己感兴趣的物品。

值得一提的是，亚马逊在做推荐时，设计和用户体验也做得特别独到。

亚马逊利用其大量历史数据的优势，量化推荐原因。

基于社会化的推荐，亚马逊会给出事实的数据，让用户信服。例如，购买此物品的用户百分之多少也购买了那个物品。

基于物品本身的推荐，亚马逊也会列出推荐的理由。例如，因为你的购物框中有“某某”，或者因为你购买过“某某”东西，所以给你推荐类似的“某某某”。

另外，亚马逊的很多推荐是基于用户资料档案计算出来的，用户的资料档案中记录了用户在亚马逊上的行为，包括看了哪些物品、买了哪些物品、收藏夹和购物车里的物品等等。当然亚马逊里还集成了评分等其他的用户反馈方式，它们都是资料档案的一部分。同时，亚马逊提供了让用户自主管理自己资料档案的功能，通过这种方式用户可以更明确地告诉推荐引擎他的品味和意图是什么。

多维度“数据资产”的争夺是行为广告的主战场

谷歌仅仅凭借“搜索数据”就成为一方霸主。事实上，大量的数据资产的价值，并没有被充分利用。谷歌尚有两类数据欲待染指而不可得：第一类，电信运营商业务数据；第二类，大型网站的数据，如 Facebook 的数据、亚马逊的数据等等。从竞争角度而言，电信运营商和大型商业网站不可能向谷歌公开数据。谷歌正在一步步蚕食传统电信运营商的地盘，电信运营商最猛烈的反击手段就是进入谷歌的腹地，在内容广告市场抢夺谷歌的利润，遗憾的是电信运营商未必有足够灵活的机制和充足的人才。

而 Facebook 网站有 10 亿活跃用户，本身就是巨大的广告平台，正在磨刀霍霍抢夺内容广告市场的第一把交椅位置。亚马逊是谷歌广告最大的金主之一，也是众多互联网广告的目的地，亚马逊自己的站内推荐系统已经是最优秀的广告平台之一。

在上一节，提到电信运营商的“数据资产”，可实现定位人群的精准性、实时性，在移动互联时代尤其如此。因此谁能利用谷歌不可能获得的“数据资产”，谁就将在这场传媒大战中获得战略优势。

“生态系统”的培育是问鼎行为广告的基石

内容广告的产业链较长，生态系统也比较复杂，回顾内容广告产业链形成过程，有助于理解本节主题。自 2005 年开始，内容广告产业链的改变开始了，一系列改变原有内容广告售卖的商业方式及技术开始出现，在线内容广告经过七年的创新，创造了媒体买方和卖方的整个新市场。

第一阶段，最开始的内容网络比较少，媒体可以直接将广告资源售卖给广告客户，如图 4-12 所示。

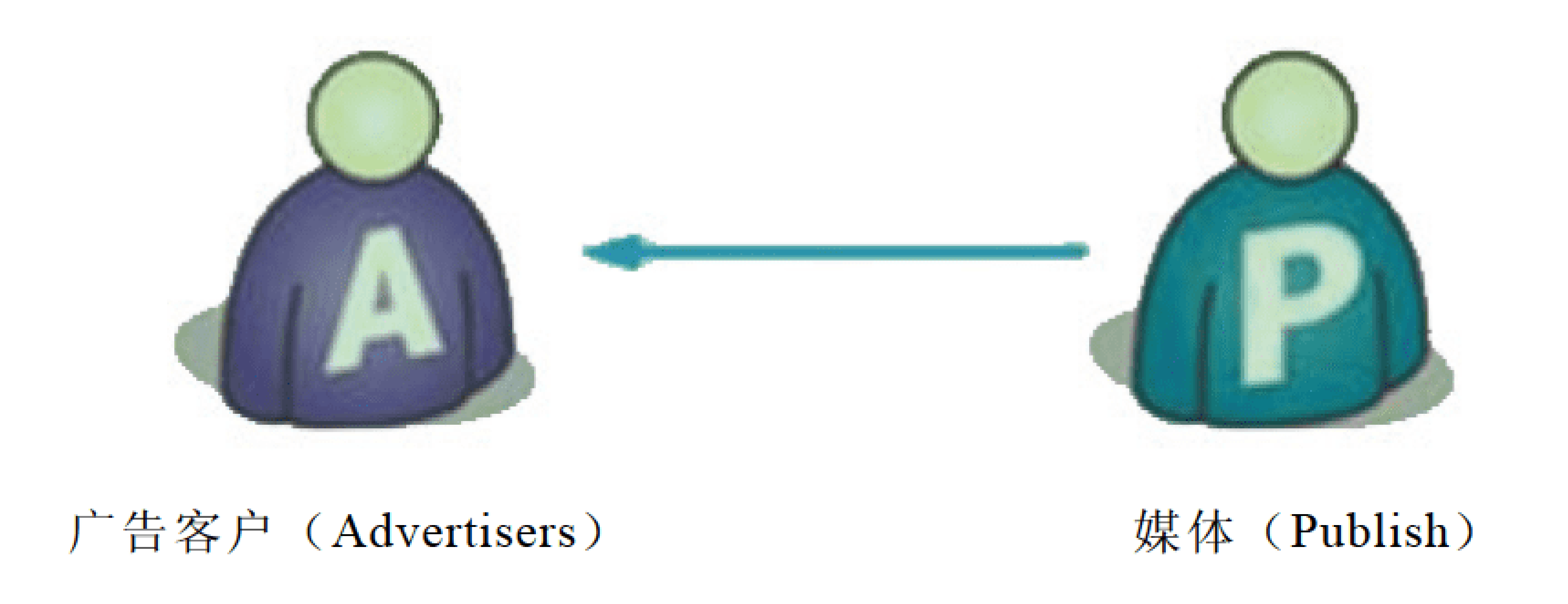


图 4-12 简单的广告交易

第二阶段，由于广告客户数和媒体数的增多，进行广告效果监控的数据也不统一，作为中介的广告网络出现，如图 4-13 所示。广告网络的出现使得广告客户和各家媒体的交易成本下降，使得不同广告客户的广告可以出现在不同的媒体上，实现一定的投放策略和竞价策略。

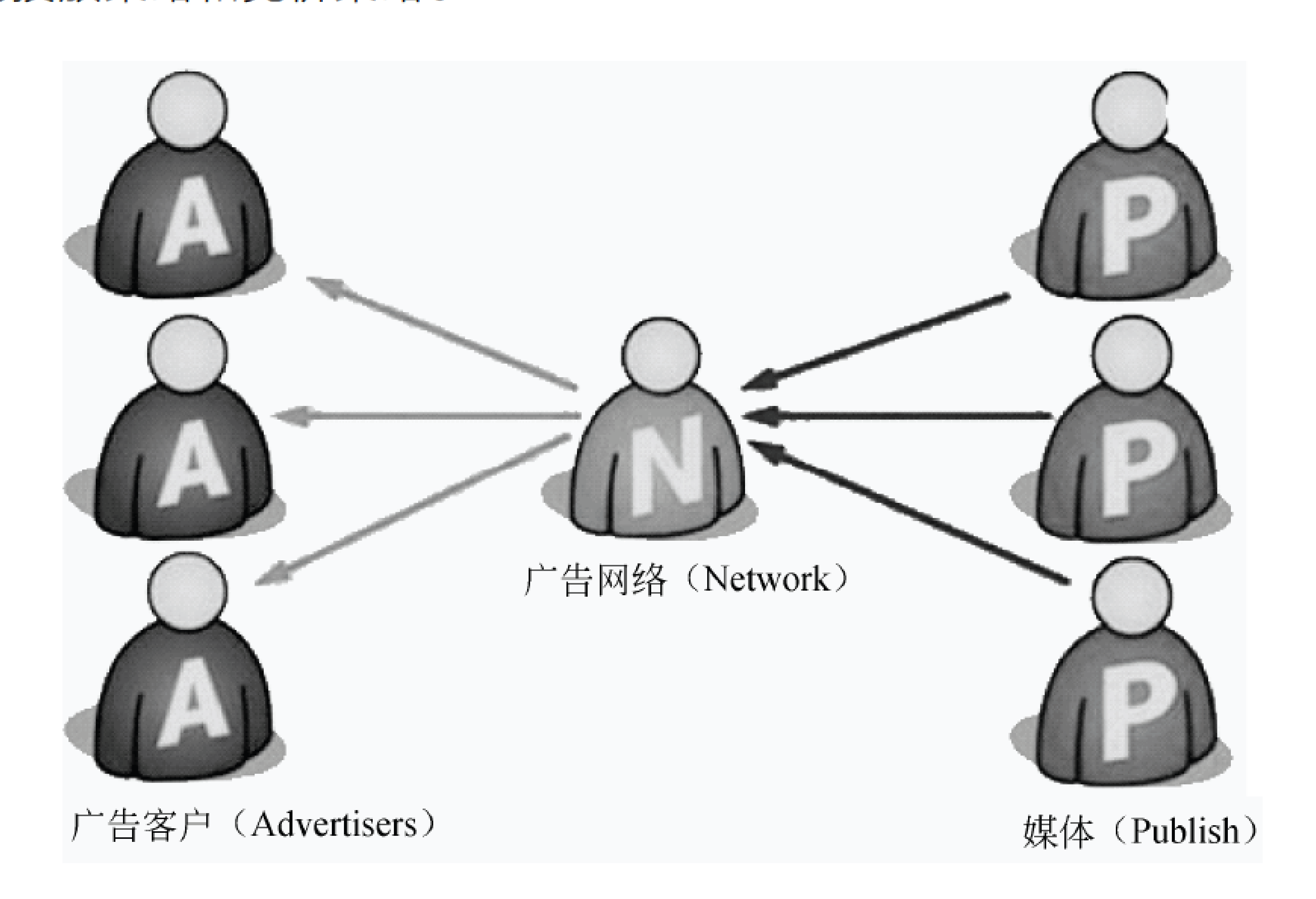


图 4-13 广告网络出现，形成广告中介平台

第三阶段，由于广告客户数、媒体数的增多，广告市场上已经不可能出现一家广告网络可以充当所有广告客户和媒体的中介，于是多家广告网络出现。不同的广

告客户如果需要投放在尽可能多的媒体上，需要跟多家广告网络打交道，而覆盖广告媒体规模的扩大也能提升广告定向投放的变现效率，为了进一步降低交易成本和提高广告定向投放的变现效率，广告交易场所（Exchange）模式诞生，并形成了以广告客户和代表广告客户利益的广告资源需求方，同时也形成了媒体和代表媒体利益的广告资源供给方，如图 4-14 所示。

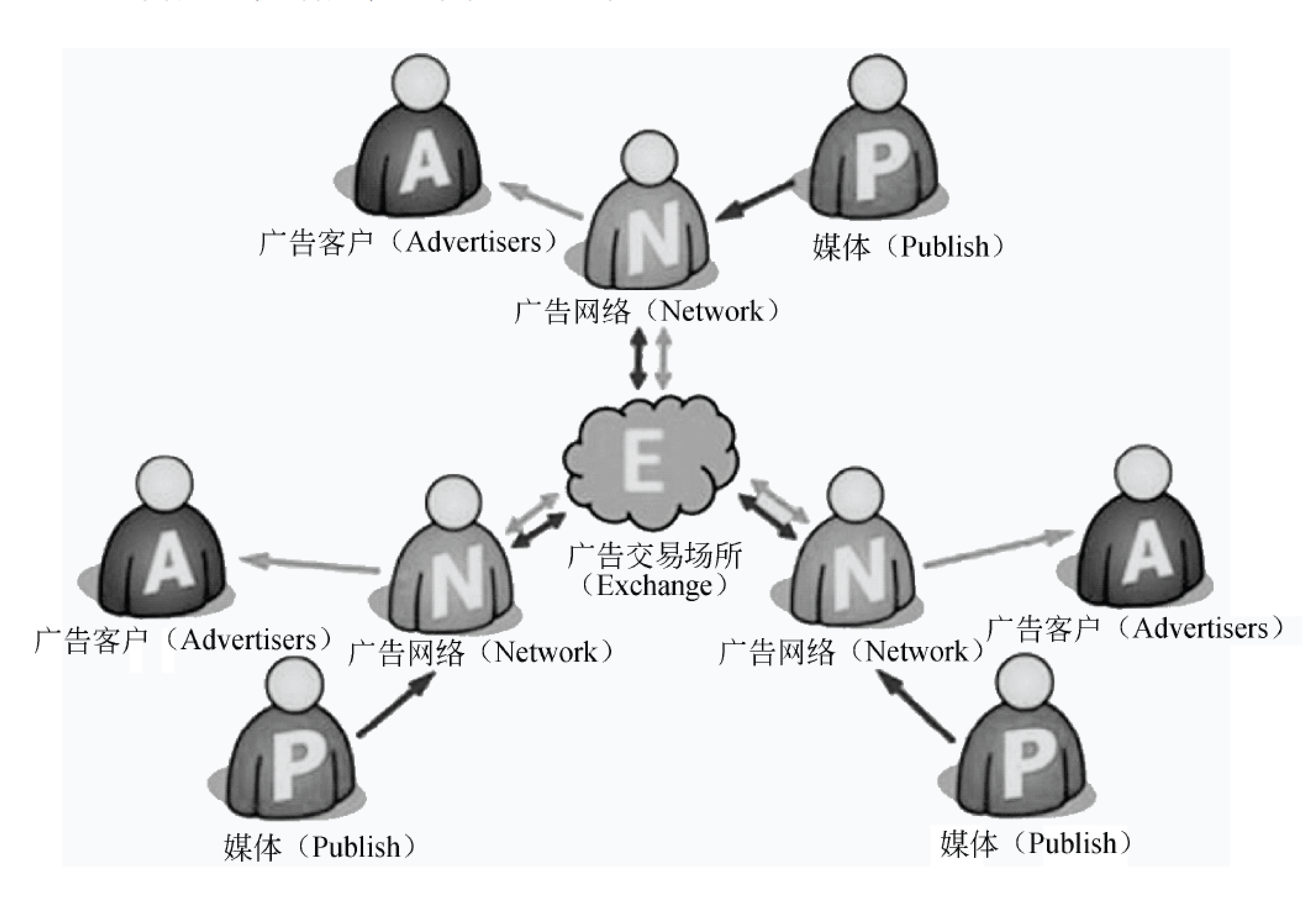


图 4-14　类证券交易所模式形成

随着市场领导者如谷歌、Facebook、雅虎投资于广告定向投放策略，广告定向效果的提升、消费者参与及新的广告创意形式共同作用加速了市场的发展，内容广告市场在重构，原有的内容广告 Network 在重塑，其他新的参与者和技术也进入 Exchange，DSP[①]、SSP[②]及 RTB 等市场，每一个部分参与者都解决了广告市场及媒体的某个特定的需求，可利用的在线数据的繁荣可以统一媒体广告资源，通过

① DSP，Demand Side Platform，需求方平台。
② SSP，Supply Side Platform，供给方平台。

创造“大部分平台一起合作的广告产业生态系统方式”来达到广告系列目标，这个广告产业生态系统的方式即与广告数据交易（Data Exchanges）平台及实时竞价(RTB)一起工作的广告网络(Ad Network)，广告市场必须创新发展以充分利用购买媒体广告资源，理解生态链各个部分的参与者以及它们是如何相互影响的是实现成功在线广告过程的关键。

图 4-15 是内容网络发展至第三阶段后的产业链概念图，这里简要介绍其中主要的三部分：Exchange、DSP 和 SSP。DSP 代表广告客户的利益，实现广告客户的利益最大化；SSP 代表媒体广告资源的利益，实现媒体价值的最大化；而 Exchange 则是中性的广告交易平台，不同代表广告客户利益的需求方平台和代表媒体资源利益的供给方平台都通过 Exchange 对每一次广告展现机会进行实时竞价，每一次广告展现机会的价值不同是因为当次展现情况下，用户属性、内容广告所在的媒体属性及当时的环境（如时间、地点等）不同，造成该次展现机会对不同广告客户的价值也不同。整个广告请求开始至返回合适广告的过程，虽然经历不同的广告产业链的各个部分和不同公司，经历大量的数据流动和计算，但是整个过程必须在 100 毫秒甚至更短的时间内完成，通过实时竞价来决定最后内容广告的过程，使得每次内容广告价值达到最大化。这样，广告客户和媒体的利益都得到了提升和优化。

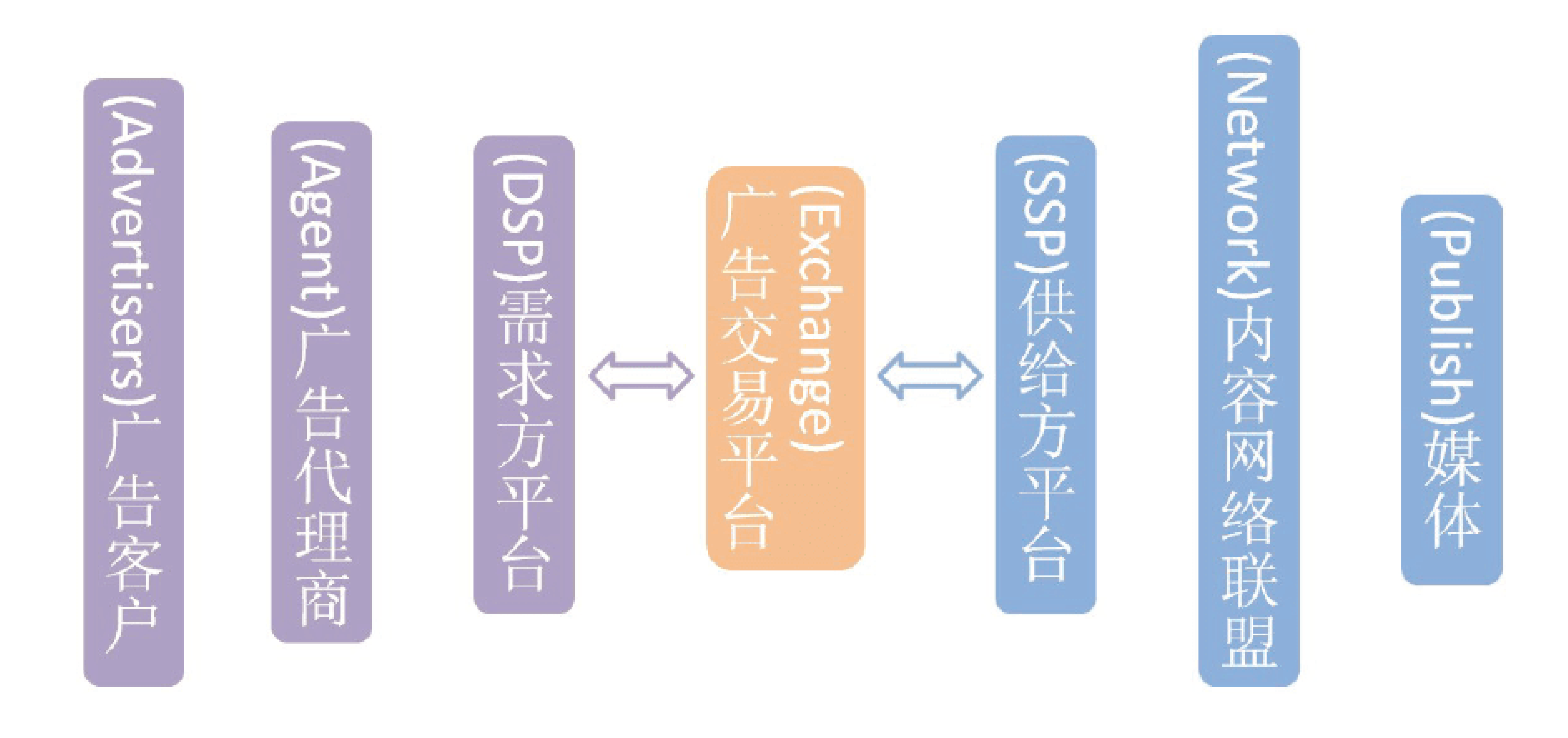


图 4-15　内容广告产业链概念图

目前，谷歌、Facebook、雅虎、微软都形成了自己的内容广告“生态环境”，新进入者必须具备更强大的技术、更丰富的数据资产、更精准的广告预测，才有可能建立新的生态环境，影响产业格局。好在内容广告技术市场变化非常迅速，尚未形成既定的竞争格局。

“大数据技术”是行为广告决胜的关键

行为广告对于大数据技术的要求非常迫切。如果没有对海量数据的存储、快速检索、深度挖掘的能力，就算所有互联网数据堆在你面前，也是无济于事。这方面需要认真地向谷歌学习，不仅学习它的商业模式，更是学习它如何凝聚人才，投入巨资，研发出许许多多引领着信息科技发展方向的东西。

2000 年后的十年，整个信息产业的技术进步落后于以谷歌为代表的大型互联网公司。无论是云计算还是大数据，都是发端于谷歌或者亚马逊，与传统的信息巨头无关。更加令人尴尬的是，谷歌也好、亚马逊也罢，都没有采用主流的商业信息系统，它们都是利用开源技术自行搭建平台。

2012 年处于风口浪尖上的京东商城，虽然从 Oracle 聘请了 CTO，但是京东的系统架构，却是从同年起开始去 Oracle 化，开始采用开源技术，并且效果很好，顶住了双十一、双十二等购物狂欢节。

第四节　大数据驱动的精准营销

提要：

1. 奥巴马再次成功当选总统，大数据技术居功至伟，甚至有人宣称政治领域的大数据时代已经到来。
2. 精准营销并不是新鲜的概念，但是如果叠加大数据技术，那么引起变革的并不仅仅限于营销领域，而是完全颠覆了企业的经营模式。

广告只不过是营销的一种手段，在本章最后一节，来讨论大数据对营销各个环节的影响。

没有人比美国总统更擅长营销。罗斯福总统上任之初，碰到全球性的经济危机，银行成批的倒闭，挤兑风潮遍及全国。就在罗斯福宣布就职的那一天，美国金融的心脏停止跳动，证券交易所正式关闭。罗斯福临危受命，在温暖的壁炉旁接受电台记者采访，发表亲切、随意的讲话，阐述经济政策，鼓舞民众的士气。罗斯福在 12 年的总统任期中，共做了 30 次“炉边谈话”。总统亲切、自然的语调通过无线电波，传播到围坐在收音机旁的每一个家庭。

“是在肯尼迪任期内和因为他，电视成为确定一位总统领导国家能力的一个重要的——也许是至关重要的决定因素”[①]。肯尼迪是第一位名副其实的“电视总统”，在此之后，电视辩论成为左右美国总统大选的一个重要环节。

无论是收音机还是电视机，传播都是单向的。互联网的诞生，使沟通和传播的手段得到质的飞跃。对选民数据的收集和挖掘，使总统大选更具针对性和个性化。奥巴马成功驾驭了这个革命性的新媒体，问鼎总统宝座。

政治营销的变革——大数据总统奥巴马

2012 年 3 月，奥巴马亲自宣布美国政府的《大数据研究与发展计划》。2012 年 11 月，奥巴马依靠大数据技术再次当选美国总统。2008 年、2012 年奥巴马的两次胜选，都与其背后的数据分析团队密不可分，数据分析的工作始终贯穿奥巴马竞选的全过程，包括获取有效选民、资金筹集、有效分配竞选资源和竞选结果预测等，其发挥了巨大的作用。这让人不禁感慨，不知是大数据成就了奥巴马，还是奥巴马成就了大数据。

大数据在总统选战的三个至关重要的环节，都发挥了难以替代的作用：帮助筹

---

① 布劳尔　《约翰•F•肯尼迪》，第 119 页

集竞选资金、分配竞选资源、预测大选结果。

奥巴马背后的数据分析团队在 2008 年奥巴马首次竞选美国总统时就已存在并发挥作用。而在 2012 年，数据分析团队的工作人员五倍于上届的规模，且进行了更大规模与深入的数据挖掘工作。奥巴马的数据分析团队要用数据去衡量这场竞选活动中的每一件事情。在政治活动中运用数据分析的目的，在于充分利用在竞选中可获得的选民资料、行为、支持偏向等多方面的大量数据。数据分析团队试图挖掘这一连串数据并预计出选民的选举模式，这将使奥巴马竞选团队的筹集资金和花费都更加精确和有效率，比如通过利用海量数据的分析来帮助奥巴马筹集到 10 亿美元竞选资金，如何重新制订了电视广告投放，如何做出“摇摆州”选民的详细模型以提升电话、上门投递邮件、社会化媒体等手段的利用效率。

筹集竞选资金

奥巴马的数据分析团队帮助奥巴马筹集到了超预期的 10 亿美元竞选资金。首先，数据分析团队将民调专家、筹款人、选战员工、消费者数据库等所有获取到的数据都聚合到一块。这个组合起来的巨大数据库不仅能够帮助竞选团队发现选民并获取他们的注意，还能获知哪些类型的人有可能被某种特定的事情所打动或说服；得到这些可能被说服的内容后，还将帮助竞选团队按选民最重要的优先诉求来排序。数据分析团队通过将选民的消费者数据建模还能帮助预测哪些人会在网上捐钱。新的大数据库能让竞选团队筹集到比他们预料到的更多的资金。

奥巴马通过网上筹集到的资金中的极大部分是通过一个复杂的、以度量驱动的电邮营销活动而来。在这里，数据收集与分析变得异常重要，采用了不同的标题、发送者与信息内容以筹集更多的竞选资金。

通过数据分析发现，注册了“快速捐献”计划[1]的人，捐出的资金是其他捐献者

[1] “快速捐献”计划允许在网上或者通过短信重复捐钱，而无须重新输入信用卡信息。

的四倍。所以该计划被拓展开来，然后以物质刺激加以激励。在 2012 年 10 月底时，该计划是竞选团队对支持者传递信息的重要组成部分。

数据分析团队通过对支持者数据的收集、分析和挖掘后，发现支持者喜欢竞赛、小型宴会和名人，于是奥巴马竞选团队创建了与奥巴马共进晚餐的“帕克竞标”来为奥巴马筹集竞选资金。

事实上，奥巴马募集到的资金尽管与对手罗姆尼募集的资金规模不相上下，但奥巴马从普通民众直接募集到的资金是罗姆尼的近两倍，说明奥巴马受到更多普通民众的欢迎和支持。

竞选资源分配

数据分析团队会得出数据处理结果，告诉竞选团队赢得这些州的机会在哪，从而使竞选团队可以更有效地进行资源分配。

线上，动员投票的工作首次尝试大规模使用 Facebook，以达到上门访问的效果。

数据同样让竞选团队把总统送往通常在竞选阶段晚期不会出现的地方。2012 年 8 月，奥巴马决定到社会化新闻网站 Reddit 去回答问题，因为一大批奥巴马的动员目标在 Reddit 上。

数据也帮助了竞选广告的购买。奥巴马团队 2012 年的电视广告购买效率比 2008 年提高了 14%，这确保奥巴马竞选团队通过广告与其可劝服的选民对话。

预测竞选产出

奥巴马数据分析团队用了四组民调数据，建立了一个关键州的详细图谱。分析团队做了俄亥俄州 29000 人的民调，占了该州全部选民的 0.5%，使数据分析团队深入分析特定人口、地区组织在任何给定时刻里的趋势。这是一个巨大的优势：当第一次辩论后民意开始滑落的时候，他们可以去看哪些选民改换了立场，而哪些

没有。

民调数据与选民联系人数据每晚都在所有能想象到的场景下被计算机处理、处理、再处理。“我们每天晚上都在运行 66000 次选举。”计算机模拟竞选，用以推算出奥巴马在每个“摇摆州”的胜算。

数据驱动的决策对奥巴马——美国历史上的第 44 位总统的续任起到了巨大作用，也是研究美国 2012 大选中的一个关键元素。这同时也是一个信号，表明华盛顿那些基于直觉与经验决策的竞选人士的优势在急剧下降，取而代之的是数据分析专家与计算机程序员的工作，他们可以在大数据中获取信息，洞察选举形势。在政治领域，大数据的时代已经到来。

营销领域的变革——F2C

企业营销虽然没有总统选战这样万众瞩目，但如果不向总统们取经，显然是要被时代抛弃的。就营销而言，没有大数据就像没有预警飞机的战斗编队，几乎没有抗打击的能力。在大数据时代，企业营销如果不能有效地针对单个消费者或消费群体来进行个性化营销，营销的价值就需要重新讨论和判断。

IBM 公司曾经采访了超过 1 700 位首席营销官（CMO），推出了一份以营销转型推进中国企业成长的报告，其中提到 13 项变革因素。这些 CMO 投票选择的前五项变革因素是：第一，数据爆炸；第二，渠道和设备选择的增加；第三，不断变化的消费者特征；第四，高速增长的市场机遇；第五，品牌忠诚度的下降。当让这些 CMO 选择企业有哪些因素尚未做好准备时，结果还是这五项。可见，大数据将影响营销领域变革的方方面面。

F2C 是亿赞普公司提出的新型营销理念，即帮助企业直接把产品从工厂（Factory）传递到消费者（Consumer）手中，工厂是全球的，消费者也是全球的。整个模式可以理解为“前店后厂”的模式，只不过“店”和“厂”不是地域连接的，而是通过创新的媒体和成熟的互联网商业模式相结合而实现的。

基于大数据的飞利浦全系列产品的全国网络传播

飞利浦作为世界最大的电子公司之一，其产品辐射各个电器类目。作为另一个“极有潜力的本土市场”，中国市场成为飞利浦整个营销战略上的重点地区之一。因此，负责小家电的飞利浦优质生活事业部大中华区与欧洲、美洲等市场地位并列，成为构建“商务组织”的四大核心市场。

飞利浦在剃须刀等个人护理小家电，以及榨汁机、吸尘器和空气净化器等生活小家电上都具有优势。不过，面对主要竞争对手美的在生活小家电领域的全线渗透，飞利浦精品小家电的地位受到冲击。自 2011 年 5 月起，飞利浦生活小家电和精品小家电面向全国重点直辖市和省份开展为期一个季度的推广促销风暴，扩大品牌及产品在网络上的曝光量。

在项目执行期间，传播活动需要对飞利浦空气净化器、风景时尚灯、吸尘器、剃须刀、soudbar、avent 六大产品进行品牌形象推广和促销。因此整个传播活动的目的是在扩大品牌及产品网络曝光量的同时，加深目标消费群对飞利浦各相关产品的认知，进而促进相应商业效益的增加。

中国互联网环境具有各网站数据割裂，碎片化十分显著的特征。不仅如此，飞利浦整个产品线品类繁多，传播任务繁重，传统互联网单一的品牌互动方式让飞利浦难以掌控，更不易于有效地传达信息，展示品牌形象。而飞利浦的目标是，通过整合营销策略让中国的消费者能够轻松愉悦地感受到飞利浦品牌的特性，最大化提升有限预算的 ROI。

事实上，从互联网传播角度来看，此次飞利浦广告投放面临四个大的挑战。挑战一：通过大数据的洞察，快速洞察人群和精品人群的网络行为特征和心理特征，为网络传播提供策略依据。挑战二：基于区域的销售策略来制定媒体传播策略，让线下销售与网络推广有效衔接。挑战三：解决多产品同步广告推送问题，提高有限媒体版位的利用效率；通过技术手段，让广告版位基于受众兴趣展示广告，不同的

受众看到不同的广告，提升广告版位价值。挑战四：基于亿赞普 ID 数据，实现跨区域广告频次调度，让有限广告位的销售价值最大化。同样的预算，UV 覆盖量达到常规广告投放方式的 2 倍以上。

确立传播策略，首先需要基于大数据建立数据模型，然后做出飞利浦人群的相关数据。那么，大数据是如何运用到飞利浦的全国推广中的呢？

在飞利浦项目的执行中，以海量数据存储系统为基础，通过数据挖掘和人工智能算法，对海量互联网用户、内容和相关行为进行分析，挖掘出其中蕴含的营销机会，以达到最具价值和效率的营销效果，同时获得了更高的投资回报率。

策略执行过程中，以互联网大数据分析为基础，结合互动策略、数字创意、互联网媒体采购、互联网公共关系和监测服务等进行全面的整合服务，建立了包括差异人群覆盖、品牌植入传播、多媒体组合策略、EPR 互动口碑传播以及 CRM 用户持续管理系统等一系列完善的体系，通过技术与媒体的数据化结合，形成了基于智能化投放的 361° 传播策略，全面覆盖飞利浦的目标受众，如图 4-16 所示。



图 4-16　飞利浦案例中的传播模型

这一切，已不仅仅是营销，而是需要数据、技术与营销的完美融合。

导读：

1. 2013 年是互联网金融的元年，互联网金融具有广阔的发展前景，已经初步呈现出商业生态体系，包括第三方支付、移动支付、P2P 外汇、金融渠道、网络供应链金融、网络小额信贷、P2P 网络借贷、众筹、比特币等模式。
2. 金融成为互联网第四波变现浪潮，以阿里巴巴为代表的互联网企业正在不断跨界侵入传统金融领域。
3. 互联网金融的兴起具有深刻的历史背景，是技术、政策、需求以及资本多种因素共同作用的结果。
4. 互联网金融呈现三大发展趋势：衍生金融需求，创新金融模式，重构金融格局。
5. C2C（Copy to China）已经过时，中国在这一轮变革中将会实现弯道超车，引领全球互联网金融的发展。

第五章

互联网金融

中国金融行业特别是银行业，服务了 20%客户，却有 80%企业没被服务。

——马云

2013 年可以称为互联网金融的元年。虽然互联网金融概念[①]是在 2012 年由中国投资有限责任公司副总经理谢平教授首次提出，但在 2013 年互联网金融概念的热度急速攀升。6 月 17 日，阿里巴巴推出“余额宝”产品，该产品上线 6 天用户数就突破 100 万，上线 18 天累计用户数达到 251.56 万，累计转入资金达到 66.01 亿元，创造了令金融界震惊的奇迹，随后不仅出现了“活期宝”、“现金宝”等众多“宝”类产品，而且也拉开了互联网金融的大幕。产业界、投资界、学术界乃至监管部门开始纷纷加大在互联网金融领域的布局：7 月，互联网金融千人会俱乐部成立；8 月，中国人民银行组团赴上海、深圳等地调研互联网金融并给与了支持和肯定，微信 5.0 携微信支付发布，民生电商注册成立；9 月，阿里巴巴与民生银行战略合作，上海自贸区挂牌，探索互联网金融和民营金融，习近平等领导人视察中关村，柳传志、李彦宏、雷军等企业家汇报大数据；10 月，阿里巴巴控股天弘基金，百度推出理财计划“百发”，互联网金融千人会俱乐部主办的 2013 年互联网金融峰会成功举行；11 月，首家互联网保险公司——众安在线财产保险股份有限公司揭牌。一时间，互联网金融成为社会各界争论的焦点和抢夺的重点。

互联网金融并不是一时热炒的概念，它是具有非常大的发展空间的，其中蕴含了诸多的商业模式和投资机会。就目前而言，互联网金融已经初步呈现出商业生态体系，在支付结算、资金融通和货币三大领域的营销渠道层面、产品层面、商业模式层面以及服务层面产生了很多新兴的商业模式，包括第三方支付、移动支付、P2P 外汇、金融渠道、网络供应链金融、网络小额信贷、P2P 网络借贷、众筹、比特币等。此外，针对互联网金融快速发展提供技术、数据支持的“铲子”企业也开始崛起，如图 5-1 所示。

---

① 引自谢平在中国金融四十人论坛的研究成果——《互联网金融模式研究》，报告提出可能会出现既不同于商业银行间接融资、也不同于资本市场直接融资的第三种金融融资模式，称为“互联网金融模式”。

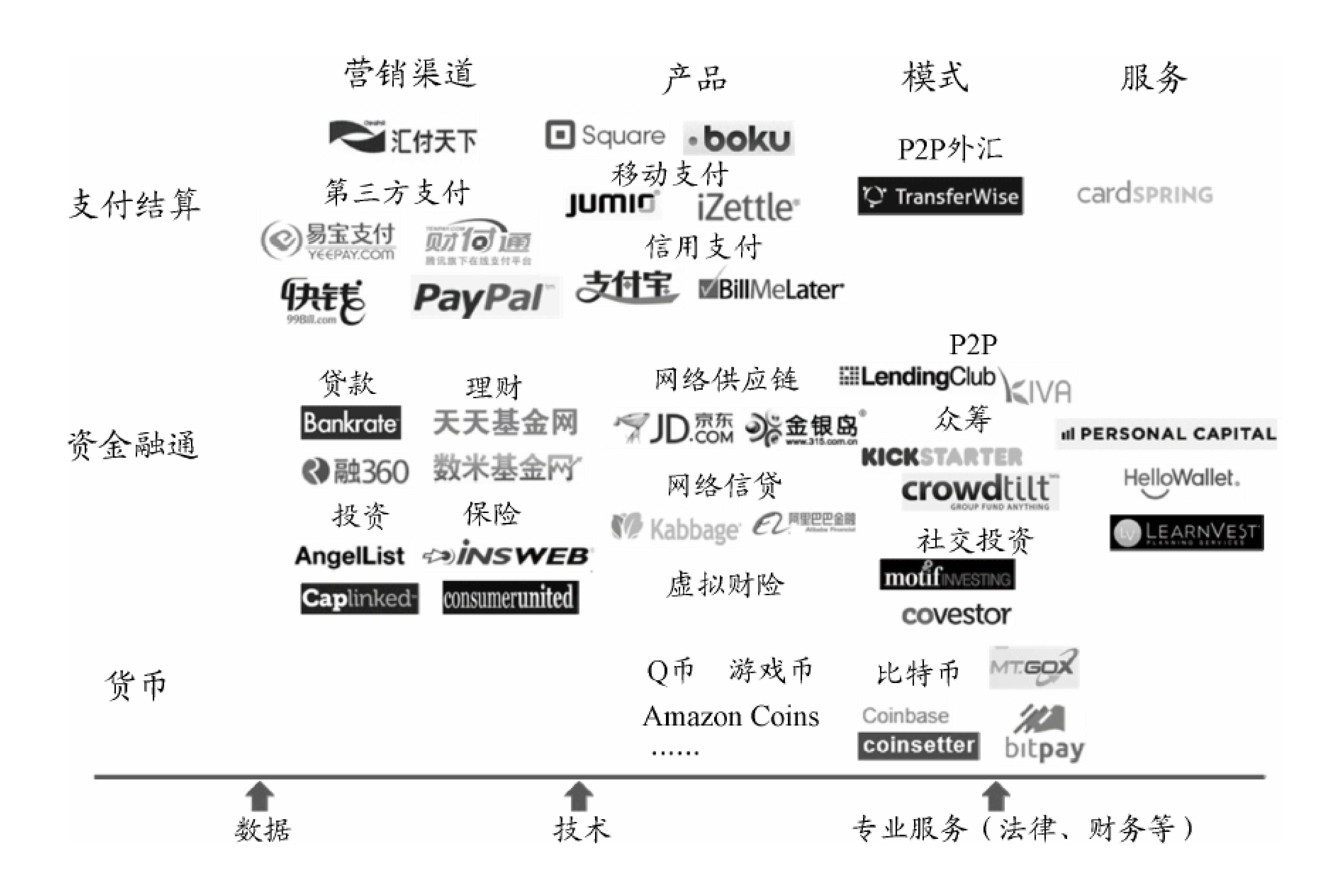


图 5-1　互联网金融产业生态地图

第一节　金融业门口的“野蛮人”掀起互联网金融浪潮

“余额宝”引爆市场

2013 年 6 月 17 日，支付宝的“余额宝”正式上线，用户可以将资金转入余额宝，在线上购买天弘基金货币基金，并从中获得较高的投资收益，如图 5-2 所示。“余额宝”是一个三方共赢的产品，为基金提供了一个快捷的渠道，为用户带来了额外收益，增强了支付宝的用户黏性，而且余额宝在货币基金的转出和消费方面进行了创新，更加方便用户使用。

“余额宝”上线 6 天，用户数就突破 100 万，上线 18 天，累计用户数达到 251.56 万，累计转入资金达到 66.01 亿元。截至 2013 年 9 月，“余额宝”资金规模已突破 500 亿元，客户数可能超过 1200 万户，令传统金融机构无不汗颜！

“余额宝”产品表面上仅仅是在“卖”货币基金的流程上进行了小小的创新，用户可以选择实时赎回，直接用于淘宝和天猫的消费，但这对于阿里巴巴以及整个市场的意义却十分重大，“余额宝”不仅为阿里巴巴进一步涉足金融做了很好的一次尝试，而且也让第三方支付等其他机构意识到新的变现机会。

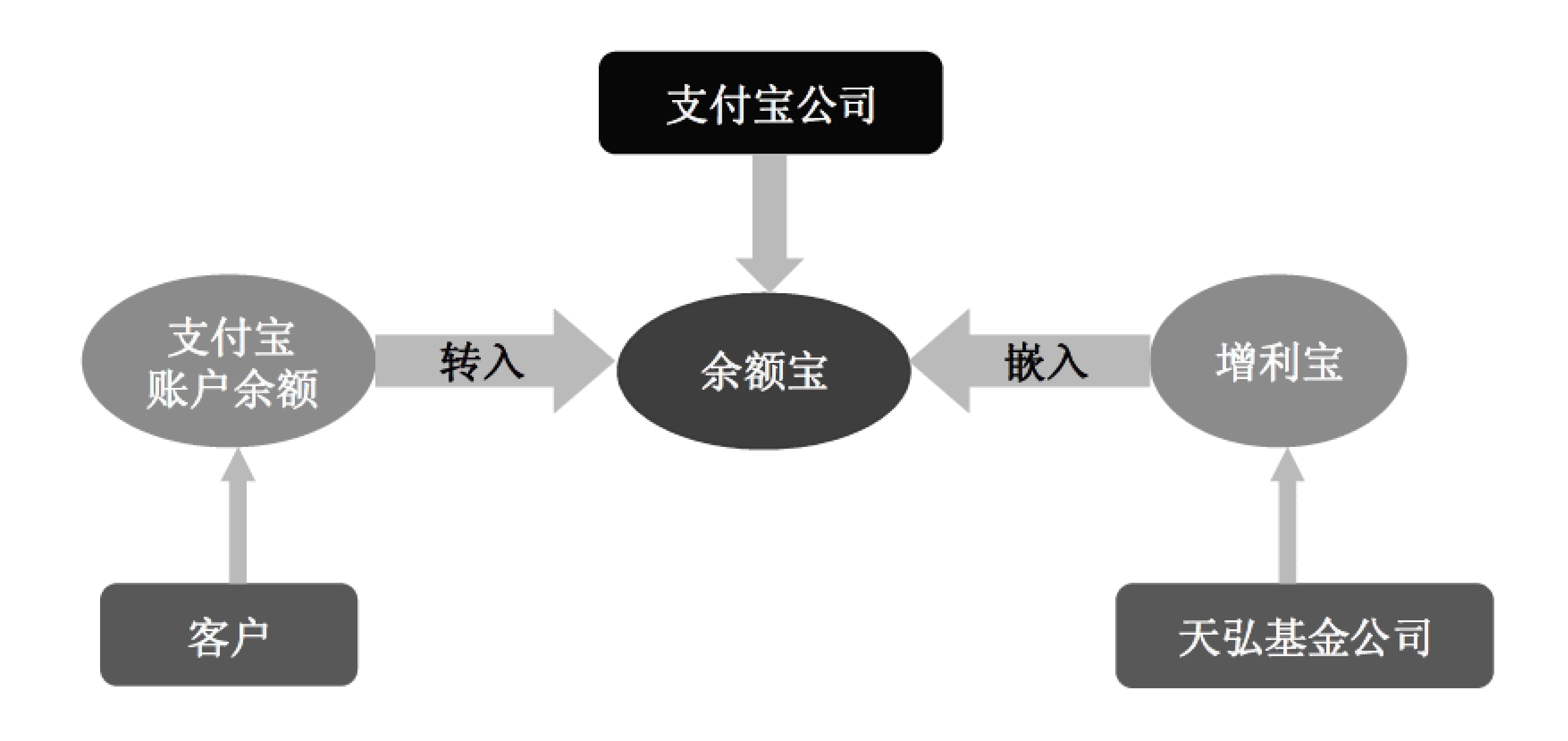


图 5-2　余额宝运作模式

一石激起千层浪，“余额宝”成功推出后，一大批“类余额宝”产品随之涌现。6 月 26 日，东方财富宣布推出针对优选货币基金的新型投资工具——“活期宝”，8 月 1 日，同花顺全资子公司浙江同花顺基金销售有限公司正式推出创新理财工具——“收益宝”，总之，市场热浪不断高涨。

阿里巴巴全面布局互联网金融

金融业是一个极其庞大的市场，包括银行、保险、证券、基金、信托等多个细分行业。以银行业为例，根据银监会的统计数据显示，2012 年我国商业银行全年累计实现净利润 1.24 万亿元，同比增长 19%，成为中国盈利最大的行业。

同时，金融是继广告、游戏、电子商务之后的互联网新型变现方式，因而也成为阿里巴巴的战略重点。2012 年 9 月，马云在网商大会上明确提出平台、金融和数据三大业务，之后，阿里巴巴在金融业务领域进行了一系列的动作和布局。2013 年 3 月，阿里巴巴集团宣布，将筹备成立阿里小微金融服务集团，负责阿里集团旗下所有面向小微企业以及消费者个人服务的金融创新业务。在组织架构方面，阿里小微金融服务集团包括支付宝共享平台事业群、支付宝国内业务事业群、支付宝国际业务事业群和阿里创新金融事业群，其中阿里创新金融事业群下面分设三大业务，分别是小贷和信用支付（包括担保公司）、保险（包括对接众安在线财产保险公司）、理财，如图 5-3 所示。

阿里巴巴已经在多个金融领域进行布局，并对传统金融体系形成冲击！借用阿里巴巴马云先生的一句话就是：“金融行业的搅局者”。支付宝对传统支付，阿里金

融对传统银行和小贷，众安在线对保险无不构成了一定的威胁，如图 5-4 所示。

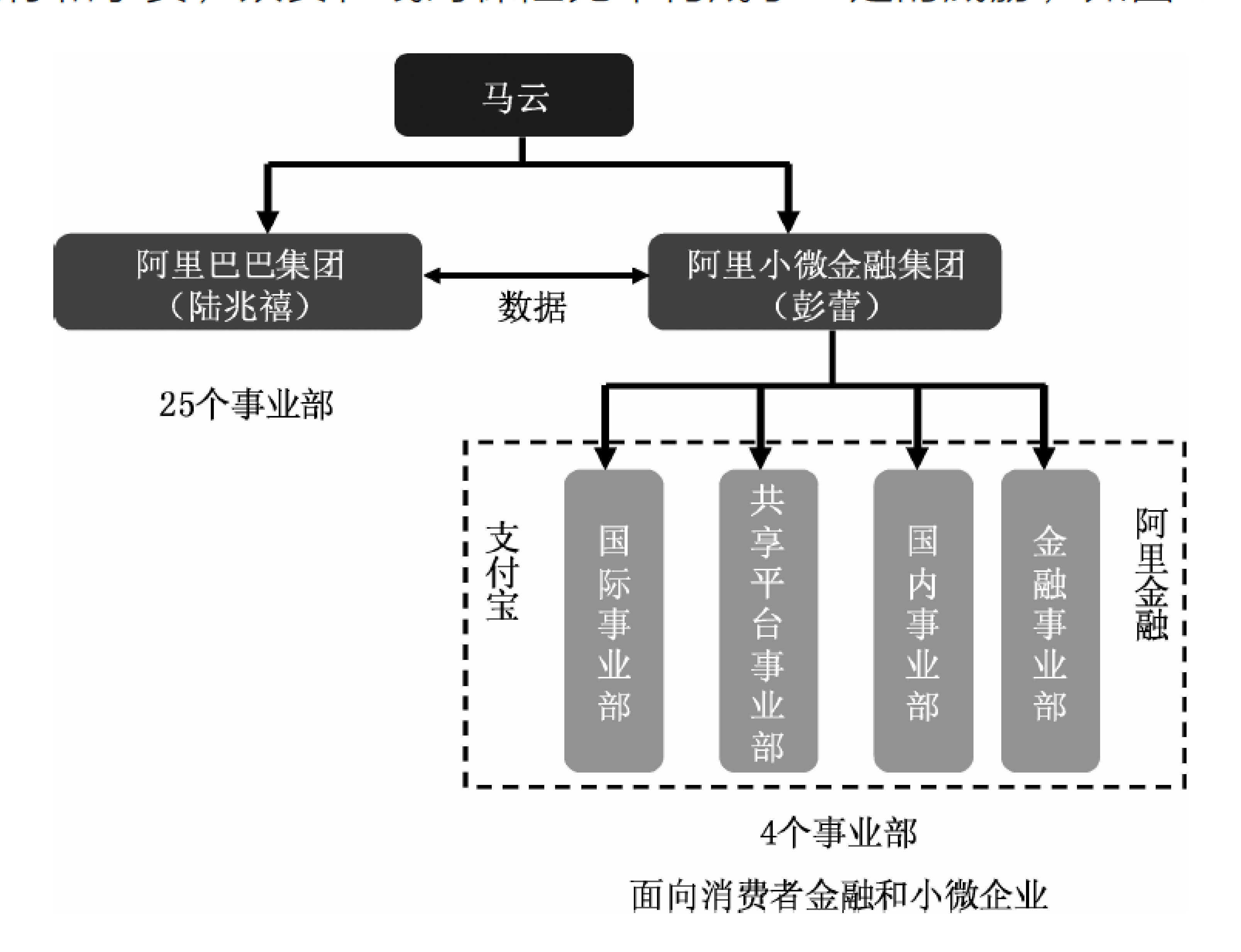


图 5-3　阿里小微金融服务集团组织架构

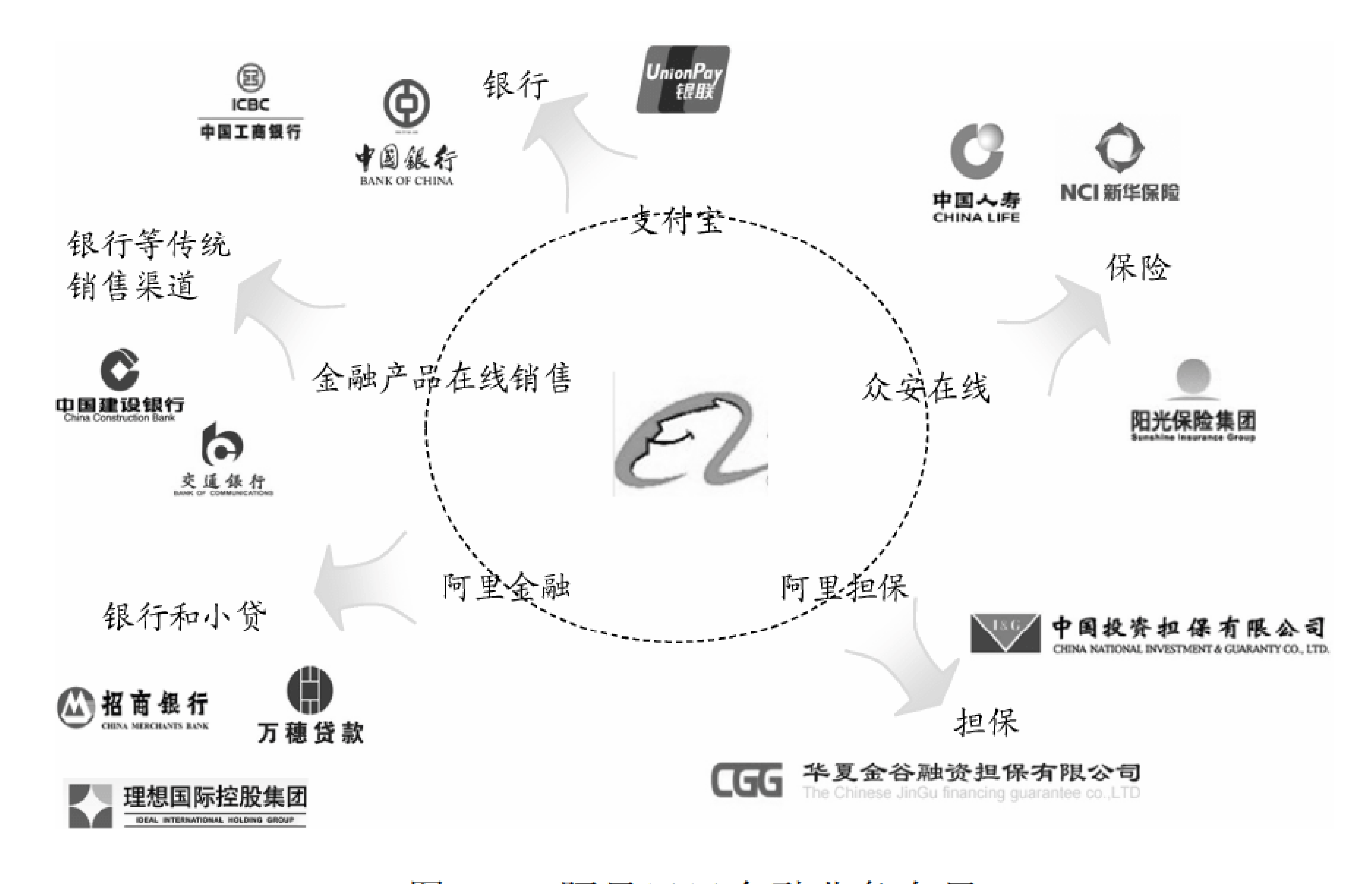


图 5-4　阿里巴巴金融业务布局

基于大数据的阿里小贷发展迅猛

在阿里巴巴的金融业务当中，阿里小贷是社会津津乐道的代表。为了解决中小企业融资难的问题，推动阿里巴巴生态体系的健康发展，阿里巴巴在 2007 年率先

与中国建设银行合作，推出融资服务，但由于各方面因素导致合作终止。随后，阿里巴巴分别于 2010 年和 2011 年联合复星集团、银泰集团和万向集团成立了浙江阿里巴巴小额贷款和重庆阿里巴巴小额贷款两家公司，开展网络贷款业务。阿里小贷业务可以分成两大类：一类是针对 B2C 平台，即为淘宝网和天猫商城的客户提供订单贷款和信用贷款；另一类是针对 B2B 平台，即为阿里巴巴中国站或中国供应商会员提供的阿里信用贷款，具体又分成循环贷和固定贷两种。阿里金融的贷款金额通常是在 100 万元以内，采用按日计息的收费方式，信用贷款和循环贷的贷款利率为 0.06%/天，其他的贷款利率为 0.05%/天，如图 5-5 所示。

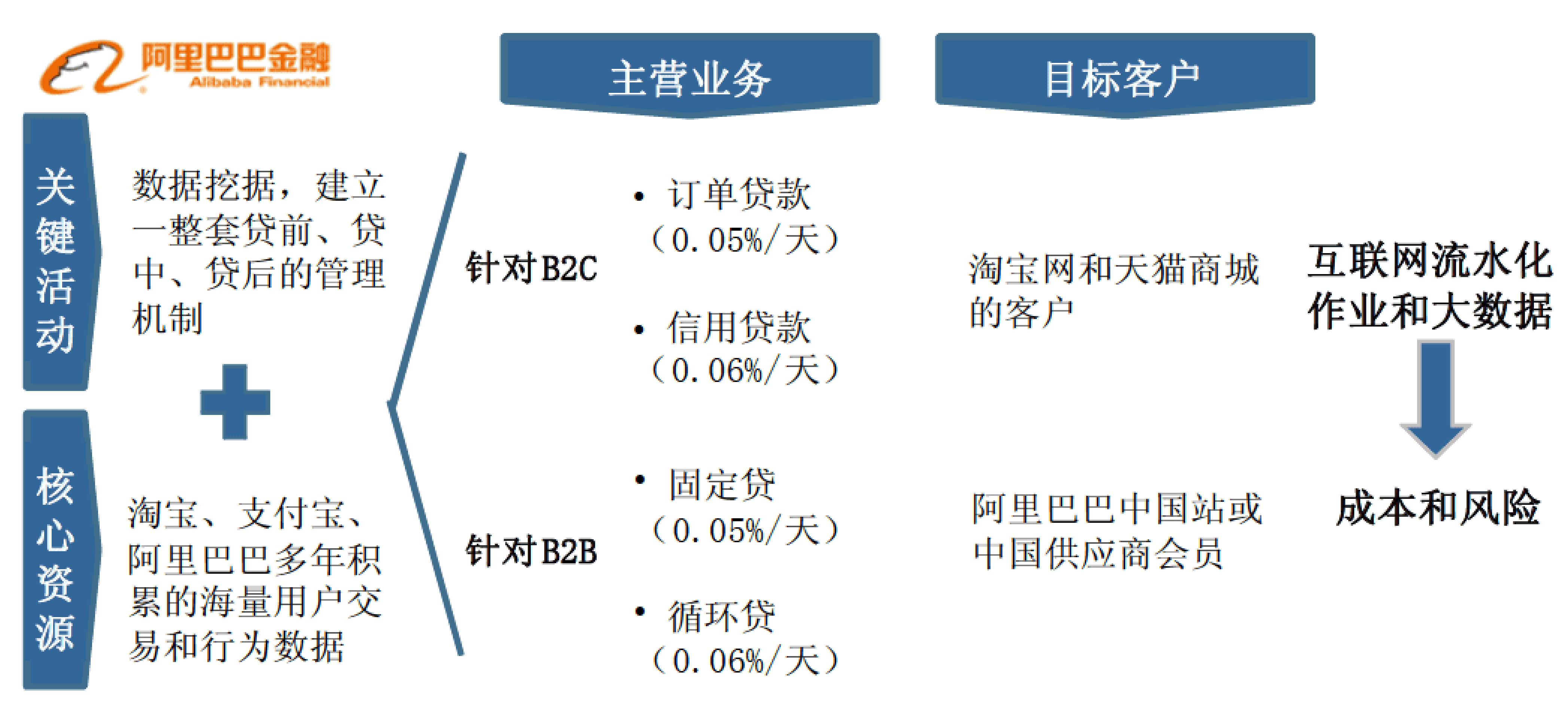


图 5-5　阿里小贷的业务模式

阿里小贷依托阿里巴巴（B2B）、淘宝、支付宝等平台多年积累的海量数据，有效地控制了小微企业的风险问题，同时借助互联网的批量化、流水化作业又大大减少了业务成本。阿里小贷凭借独特的竞争优势和商业创新，在小微企业融资领域迅速发展起来。据统计数据显示，截至 2013 年第二季度末，阿里小微贷款累计服务客户超过 32 万家，累计投放贷款超过 1000 亿元，不良贷款率为 0.87%。

在成功开展 B 端业务后，阿里金融又开始通过整合支付宝、淘宝集市和天猫商城的平台数据染指 C 端业务。2013 年 4 月，阿里金融与上海农村商业银行合作涉足“虚拟信用卡”。淘宝、天猫的用户可以获得最高 5000 元的授信，在手机支付时可以使用信用支付额度购物，由合作银行将钱支付给卖家。阿里金融通过向合作商

家收取手续费和向用户收取逾期罚息继续获利。截至目前，有 100 多万家淘宝集市卖家和近 3 万家天猫商城卖家开通了信用支付，预计该产品未来将覆盖 8000 万用户，市场潜力非常巨大。

阿里巴巴固然可怕，但更令传统金融机构担心的是，在阿里巴巴身后还有一大批互联网公司，腾讯、百度、苏宁、新浪、京东商城等企业已经推出了各自的互联网金融业务，同时 360、网易等企业也在虎视眈眈，大量“野蛮人”的入侵会在不同领域不断侵蚀传统金融机构的领土。

第二节　互联网金融爆发的历史背景

以阿里巴巴为代表的“野蛮人”在某种意义上推动了互联网金融的发展，但真正令互联网金融快速兴起的是技术、政策、需求以及资本等多种因素的共同影响，如图 5-6 所示。

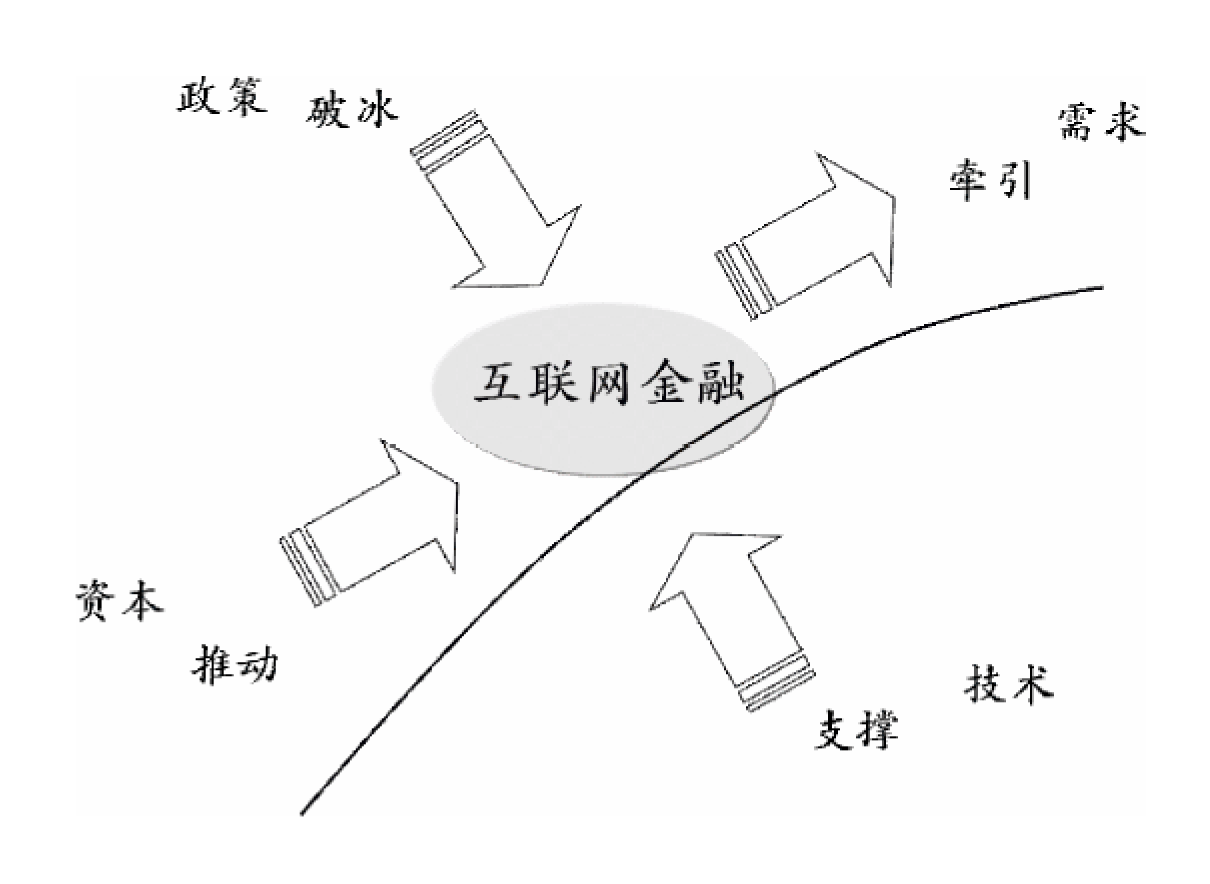


图 5-6　互联网金融兴起的历史背景

技术日新月异，提供有效支撑

金融是一个信息密集型产业，信息、技术、制度构成金融业的三大基石。纵观历史，从 19 世纪 30 年代电报的兴起，到后来的电话、计算机，乃至今天互联网、移动互联网，每一次通讯信息技术的变革都对金融业产生了巨大的影响。近些年，随着物联网、社交网络、云计算、移动互联网等新兴信息技术不断涌现，改变了传

统的信息产生、传播、加工利用的方式，信息不对称程度大幅下降，信息的获取和处理成本大幅减少，资源的配置效率大幅提升，对金融业产生了巨大的影响。

（1）搜索引擎技术通过信息搜索、组织和处理后，为用户提供检索服务，该技术满足了用户在信息大爆炸时代快速、低成本获取所需信息的需求，尤其是随着个性化搜索、情景搜索等技术的进一步发展，该功能价值将更加凸显。

（2）社交网络作为 Web 2.0 时代的关键应用，对信息处理的影响体现在三个方面：一是使得每个人不仅是信息的消费者，也是信息的生产者；二是实现了信息的双向传播，加强了互动与分享；三是将分散的信息聚合，形成一个个具有一定特征的信息节点（社区），这为信息的加工利用提供了更多的选择、更好的基础，如利用社交网络一致性特点，通过一度好友、二度好友的分析来解析该用户，由基于某个点的分析转向基于社交网络的整个面分析等。

（3）云计算技术如同物理世界中的水、电、煤一般实现了用户需求与物理、虚拟资源的动态配置，在计算机物理硬件短期内难以突破的情况下，借助分布式计算、网络存储等方式大大提高了海量数据的计算和储存能力。

（4）大数据是当前市场炙手可热的话题，联合国、美国政府、法国政府等组织都对其给予了高度重视，美国奥巴马政府甚至将其上升至国家战略高度。大数据具有规模大（Volume）、速度快（Velocity）、类型多（Variety）和价值大（Value）的 4V 特征，其不仅是适应时代发展的技术产物，更是一种全新的思维理念，即基于数据资产的商业经营模式。

（5）移动互联网在原有桌面互联网的基础上进一步打破了时间和空间对于用户的禁锢，不仅增强了信息传播的时效性，而且让用户可以随时随地进行交易、支付结算等，大大提高了金融交易的可获得性，释放了部分被束缚的需求。

金融改革稳步推进，政策环境不断破冰

金融改革和金融创新稳步推进，政策正在不断破冰。中国经济经过三十年的高速增长，而今又站在了一个关键的十字路口，经济结构调整和转型升级成为了支撑新一轮增长的关键。而金融和实体经济密不可分，为了发挥金融对未来经济发展的重要支持作用，推动金融改革和金融创新成为了新一届政府的重要着力点，仅仅在过去两个月里面，政府就出台了一系列的政策措施。6 月 19 日，国务院推出 8 大措

施助经济结构调整和转型升级，指出要推动民营资本进入金融，鼓励金融创新；7月 5 日，金融“国十条”出台，再次强调要扩大民间资本进入金融业；7 月 19 日，央行进一步推动利率市场化改革，取消贷款 0.7 的下线；8 月 12 日，国务院办公厅关于金融支持小微企业发展的实施意见……此外，监管层的相关负责人也在多个场合表态支持互联网金融的发展，这些为互联网金融的兴起提供了良好的政策环境。

需求真空大量存在，新兴用户习惯形成

巨大的市场需求空白，大量中小企业的融资需求和人们的价值增值需求未被有效满足。马云在今年 6 月的外滩金融峰会上直言中国的金融业，仅仅服务了 20%的客户，难以支撑 30 年以后的中国所需要的金融体系。这句话有一定的现实意义，中国中小企业创造了 60%的国民财富，贡献了 50%的财政税收，提供了 80%以上的城镇就业，创造了 65%的发明专利和 80%以上的新产品开发，但融资难的问题却一直束缚着中小企业的发展，与之对应的是中国人们投资渠道的缺乏，价值增值需求难以得到满足，正是这样的市场需求成为了互联网金融迅猛发展的根本驱动力。

80 后尤其是 90 后几乎是伴随着互联网成长起来的一代，这部分群体对互联网、移动互联网具有高度依赖的特点，已经养成了在网上获取信息、娱乐、购物的习惯，而且这部分群体正在逐步成为中国社会消费的中流砥柱。这已经不是一种趋势，而是一种现实。根据一项调查显示，中国人平均每天用在手机上网方面的时间是 158 分钟，远高于全球范围的平均值 117 分钟，其中 25～35 岁的 80 后是最主要的群体，每天手机花费在上网的时间更多。

80 后、90 后追求网络消费、科技消费、个性化消费，这与传统的 60 后、70 后的消费需求和消费习惯存在巨大的差异。由此，这对于传统的金融业提出了全新的要求，金融机构必须采取适合该部分群体的产品、服务内容和方式，才能适应这个时代的发展。

资本市场热炒，加速互联网金融兴起

互联网金融方兴未艾，国内外大小金融机构、电商和创投均对这一行业给予了前所未有的关注和投资，大量资金开始向互联网金融产业涌入，加速了互联网金融

的快速发展。

IDG 资本和宜信公司于 9 月 27 日在北京共同宣布发起成立“IDG · 宜信金融创新基金”，首期投资规模 1 亿美元，将主要关注相关公司的中早期阶段甚至种子期阶段。2013 年 8 月 30 日，石景山区召开国家服务业综合改革试点区互联网金融产业基地揭牌新闻发布会，宣布建立北京互联网金融产业基地。该区将每年安排 1 亿元专项资金用于支持互联网金融产业基地建设。此外，还将成立互联网金融资信公司，建立互联网金融征信平台，区政府将与首钢总公司共同设立总规模为 3 亿元的互联网金融产业投资基金，专门投资于初创期和成长期的企业。

针对互联网金融的投资案例和交易额度也在不断升高，如表 5-1 所示。融 360 成立于 2011 年 10 月，是一家融资贷款及信用卡的搜索、推荐与服务平台，为小微企业和个人消费者免费提供便捷、划算、安全的金融服务。月申请贷款额突破百亿元，覆盖城市超过 30 个。2013 年 7 月获得 3000 万美元 B 轮投资，红杉资本领投，KPCB 中国、光速创投等跟投；此前融 360 曾获得 KPCB 中国、光速、清科等合计 700 万美元 A 轮投资。

表 5-1　2013 年 1 月至 2013 年 9 月互联网金融领域创业企业的投资情况梳理

时间	公司	行业细分	投资机构	融资金额	融资轮次
2013.1	铜板街	理财	华创资本	未透露	天使投资
2013.1	比特币交易网	虚拟币/比特币	未透露	100 万美元	天使投资
2013.1	易宝网络	保险	凯辉投资;美国 FTV Capital	1000 万美元	不明确
2013.1	大家投	众筹合投	深圳创新谷	未透露	天使投资
2013.1	卡小二	信用卡	蓝驰创投	数百万美元	A 轮
2013.1	哆啦宝	支付	未透露	未透露	天使投资
2013.1	钱多支付	支付	红杉资本	数千万人民币	A 轮
2013.2	多钱网	贷款	3 家 VC 机构	数千万人民币	A 轮
2013.4	好贷网	贷款	同创伟业	千万元	A 轮
2013.5	点融网	贷款	东方资产管理公司	数千万人民币	A 轮
2013.5	杭州捷蓝信息	支付	深创投	数千万人民币	A 轮
2013.7	雪球财经	股票基金	红杉资本；晨兴创投	1000 万美元	B 轮
2013.7	盒子支付	支付	金沙江创投；国微技术	1000 万美元	B 轮

续表

时间	公司	行业细分	投资机构	融资金额	融资轮次
2013.7	融 360	贷款	红杉资本；KPCB;光速创投	3000 万美元	B 轮
2013.7	91 金融超市	综合/其他	经纬中国；宽带资本CBC	数百万美元	B 轮
2013.7	MEIX 美市网	股票基金	创业工厂	未透露	种子天使
2013.8	上海捷银支付	支付	平安集团	未透露	收购
2013.8	盈盈理财	理财	未透露	数千万美元	A 轮
2013.9	卡牛/随手记	理财	红杉资本	千万美元	A 轮
2013.9	挖财	记账理财	IDG	千万美元	A 轮

资料来源：IT 桔子

第三节 互联网金融的三大趋势

在全新的时代，金融业将发生翻天覆地的变化，总体可以表现为三个方面：衍生金融需求、创新金融模式和重构金融格局。

衍生金融需求

作为现代服务业的重要组成部分，金融主要是为实体经济服务。在信息技术的推动下，传统实体经济形态正在发生巨大变化，在经济形态的转变过程中，伴随着大量新的金融需求的衍生。

虚拟财险便是新金融需求的代表，由于虚拟经济的快速发展，网络世界对金融业提出了新需求，也是基于此，虚拟财险才有了相应的市场基础。以游戏为例，截至 2012 年年底，中国网络游戏用户规模约 3.36 亿人，虽然增速已经放缓，但已经拥有足够规模的用户基础。同时，随着移动互联网的发展，中国手机游戏用户规模开始爆发增长，截至 2012 年年底，中国手机游戏用户规模达到 1.39 亿人，较 2011 年增长了 33.2%。游戏账号、游戏中的虚拟币、物品装备都融入了用户的时间、精力、情感以及金钱，但由于网络安全问题，各种失盗事件层出不穷，这便为虚拟财产保险提供了市场机会。此外，其他虚拟币、网络支付近年都实现了大幅增长，基

于安全问题的各类虚拟财产保险必然拥有广阔的市场前景。

创新金融模式

以大数据为代表的新型技术将在两个层面改造金融运营模式：一是金融交易形式的电子化和数字化，具体表现为支付电子化、渠道网络化、信用数字化，是运营效率的提升；二是金融交易结构的变化，其中一个重要表现便是交易中介脱媒化、服务中介功能弱化，是结构效率的提升。

（1）电子支付日趋主流，无现金社会即将到来

货币形态与支付方式始终朝着低成本、高效率的方向演进。电子支付作为新兴的支付方式，流通速度更快、效率更高，省掉了币材成本，流通费用也较低，且应用更为方便，因而必然会成为未来的主流支付方式，无现金社会即将到来。

支付方式与货币形态密切相关，最早的货币是贝，古书有“夏后以玄贝”的说法，后来演化为金属铸币，因为金属具有价值较高、易于分割、易于保存和便于携带的特点，所以成为了一种很好的币材选择。但由于交易不断的扩大，世界上有限的金银等金属难以满足实际的需求，纸币开始出现。纸币最初是与金银挂钩，但随着布雷顿森林体系的崩溃，纸币与金融开始脱钩，货币真实价值开始向信用方式转变。

20 世纪中叶以后，信用卡和借记卡开始兴起，经过几十年的发展，磁卡已经集存款、消费、结算和理财等多功能于一体，应用更加方便快捷，在支付领域占据着重要的地位，而现金支付情景日益减少。为了提高安全性，扩大业务创新空间，现在磁条卡也慢慢被智能卡所取代。2011 年 3 月，中央人民银行开始全面启动银行磁条卡向 IC 卡迁移工作，在 2012 年 7 月中央人民银行又宣布，自 2013 年 1 月 1 日起，全国性商业银行均要发行金融 IC 卡。

由于互联网技术的普及和发展，电子支付方式日趋成为主流。除了银行开设网银之外，两种新兴方式显示出了非常强大的生命力，一是第三方支付，二是移动支付。

第三方支付已经成为一股重要金融力量，业务功能不断扩张，大大提升了支付结算的电子化速度，但竞争也日趋激烈。2010 年 6 月，中央人民银行发布了《非金融机构支付服务管理办法》，并于 2011 年开始颁发非金融机构支付业务许可证（简称“第三方支付牌照”），为第三方支付提供了政策支持和规范，从此第三方支付结束了野蛮生长期。截至目前，中央人民银行已经先后发放了六批牌照，共计 223 家企业拿到了

第三方支付牌照，牌照已经不是一个关键竞争资源，随着越来越多的市场参与者加入，竞争激烈程度不断提高，第三方支付公司纷纷开拓新的业务和市场，寻求差异化竞争。目前第三方支付公司的经营范围主要包括互联网支付、移动电话支付、固定电话支付、数字电视支付、预付卡发行与受理和银行卡收单等业务。其中的互联网支付在最近几年一直保持着高速增长势头，据统计，2006—2012 年中国互联网支付交易规模年复合增长速度高达 110%，2012 年的交易规模已经达到 3.82 万亿元。

移动互联网兴起的同时也带动了移动支付的快速发展，在国家政策利好、利益相关者（包括银联和银行、运营商、第三方支付、手机厂商等）大力推动的背景下，移动支付得到了快速普及和发展，2012 年的中国移动支付交易规模达到了 1511 亿元，每年的增长速度也在不断提高。随着移动支付标准的落地，中国移动支付已经到了爆发式增长的前夜。

（2）信用数据化推动金融产品创新

从抵押贷款到供应链金融再到网络信贷，服务效率在不断提升，但同时对风险控制也提出了更大的挑战，而大数据的积累和应用则是解决这一问题的关键。从金融贷款产品演变过程中，可以看出“抵押物”逐步趋于虚拟化，呈现出信用数据化和数据资产化的发展规律，如图 5-7 所示。

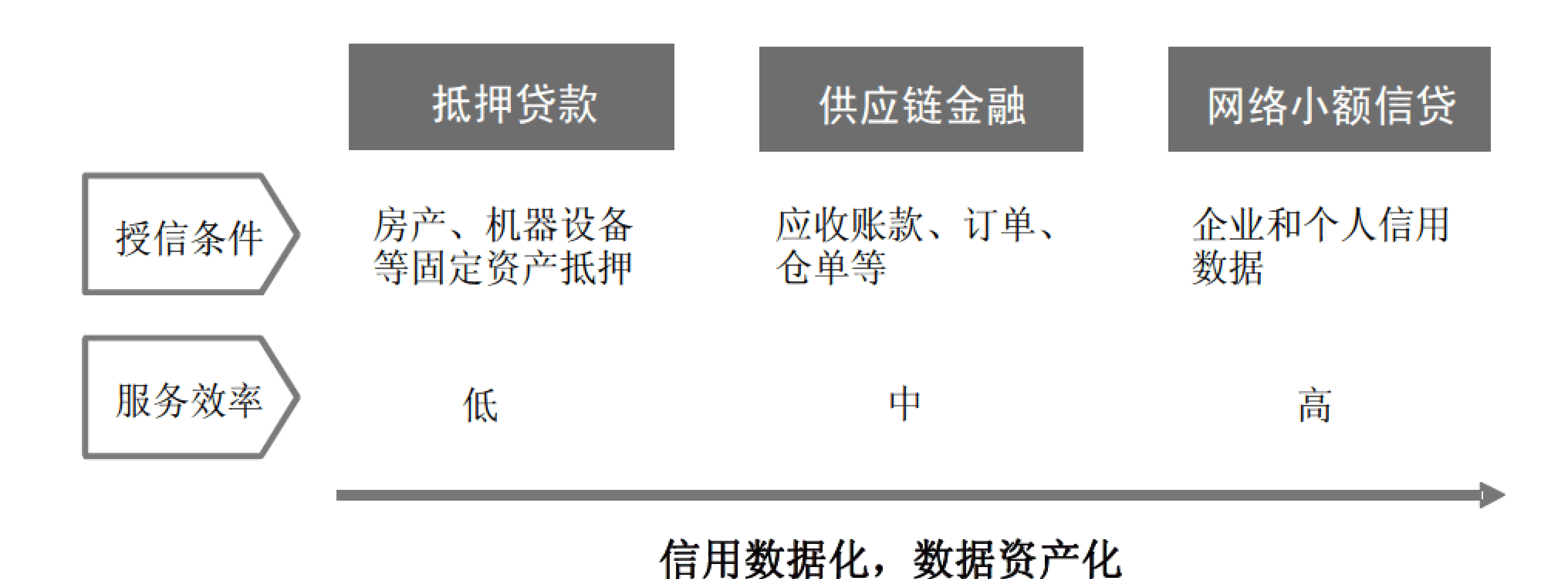


图 5-7 中国贷款产品演变分析

抵押贷款由于风险低、盈利好的特点成为中国银行等金融机构的主要业务，但弊端也很明显。中国在社会主义市场经济发展初期，进行经济金融体制改革，并全面推出抵押贷款服务，即借款人以一定的抵押品作为物品保证向银行等金融机构取得贷款，抵押品通常是建筑物、生产设备、交通工具等财产，由于其在安全性和盈

利性方面都具有明显优势，至今仍然是银行等金融机构的主要业务。但抵押贷款的弊端也是显而易见，如抵押物登记确认不统一、抵押权实现较难等，更为重要的是服务群体范围有限。

供应链金融兴起，但早期供应链金融也面临诸多问题。20 世纪末，随着物流运输行业和通讯信息技术的快速发展，全球性的业务外包活动日益增多，这在提升效率、降低成本的同时也导致了融资节点的相应增多，由于供应链各个节点参差不齐，部分节点出现资金流瓶颈并引发了“木桶短板”效应。为了解决这一问题，供应链金融随之兴起。但早期供应链金融在发展过程中面临着一系列的问题，如信息技术支持不够，中国很多银行在应收账款和预付账款等环节还需依托人工服务，这不仅降低了供应链金融的运作效率，也增加了一定的操作风险；供应链金融覆盖范围仍主要局限在重点行业和优势企业，对于中小企业的关注不够。

互联网企业将信息流、物流和资金流深度融合，开拓供应链金融业务，推动供应链金融进一步发展。2009 年，金银岛与中国建设银行、中远物流合作推出了在线融资业务——E 单通，E 单通具体又分成网络仓单融资和网络订单融资，前者是以中国建设银行认可的专业仓储公司出具的电子仓单作为质押申请融资，后者是凭借金银岛确认的电子订单向中国建设银行申请融资，其操作简单便捷，且单笔贷款最长期限为 180 天，满足了企业做行情或是短期资金周转等需求，三方共同建立了一整套服务体系和风险控制机制。2012 年 11 月，京东商城与中国银行合作推出供应链金融服务平台，为供应商提供订单融资、入库单融资、应收账款融资、委托贷款融资、协同投资信托计划和资产包转移计划等。在服务过程中，京东承担着类似中介的角色，即供应商向京东提出融资申请后，由其确认核准，并转交给银行，再由银行完成资金的发放。此外，敦煌网、苏宁等企业均推出了各自的供应链金融产品。

基于大数据的网络信贷业务崛起。目前提供网络小额信贷业务的公司主要分成两类：一是掌握大数据的互联网巨头为打造良好的生态体系而推出的网络信贷业务，如阿里巴巴的阿里小贷、亚马逊的 Amazon Lending、谷歌的广告信贷业务等；二是利用大数据开展网络信贷业务的创业公司，Kabbage 便是其中的一个典型案例。Kabbage 是一家致力于为不符合银行贷款资格的网上商家提供快速、安全的资金的信贷公司，于 2010 年 4 月上线，主要目标客户是 ebay、amazon.com、YAHOO!、Etsy、Shopify、Magento、PayPal 上的美国网商。Kabbage 通过查看网店店主的

销售、信用记录、顾客流量和评论、商品价格和存货等信息，以及其在 Facebook 和 Twitter 上与客户的互动信息，并借助数据挖掘技术（其中一个比较主要的专利技术是“为在线拍卖和市场环境提供流动资金的工具”，美国专利号 7983951），来最终确定是否为他们提供贷款以及贷款金额和贷款利率，其贷款期限最长为 6 个月，贷款月利率在 2%到 7%之间。Kabbage 用于贷款判断的支撑数据一方面来源于网上搜索和查看，另一方面则来源于网上商家的自主提供，且提供的数据多少直接影响着最终的贷款情况，同时 Kabbage 也通过与物流公司 UPS、财务管理软件公司 Intuit 合作，扩充数据来源渠道，如图 5-8 所示。Kabbage 的商业模式适应了市场发展要求，上线不到 1 年，就得到数千家商户的支持，每家商户的平均贷款资金为 1 万美元左右。

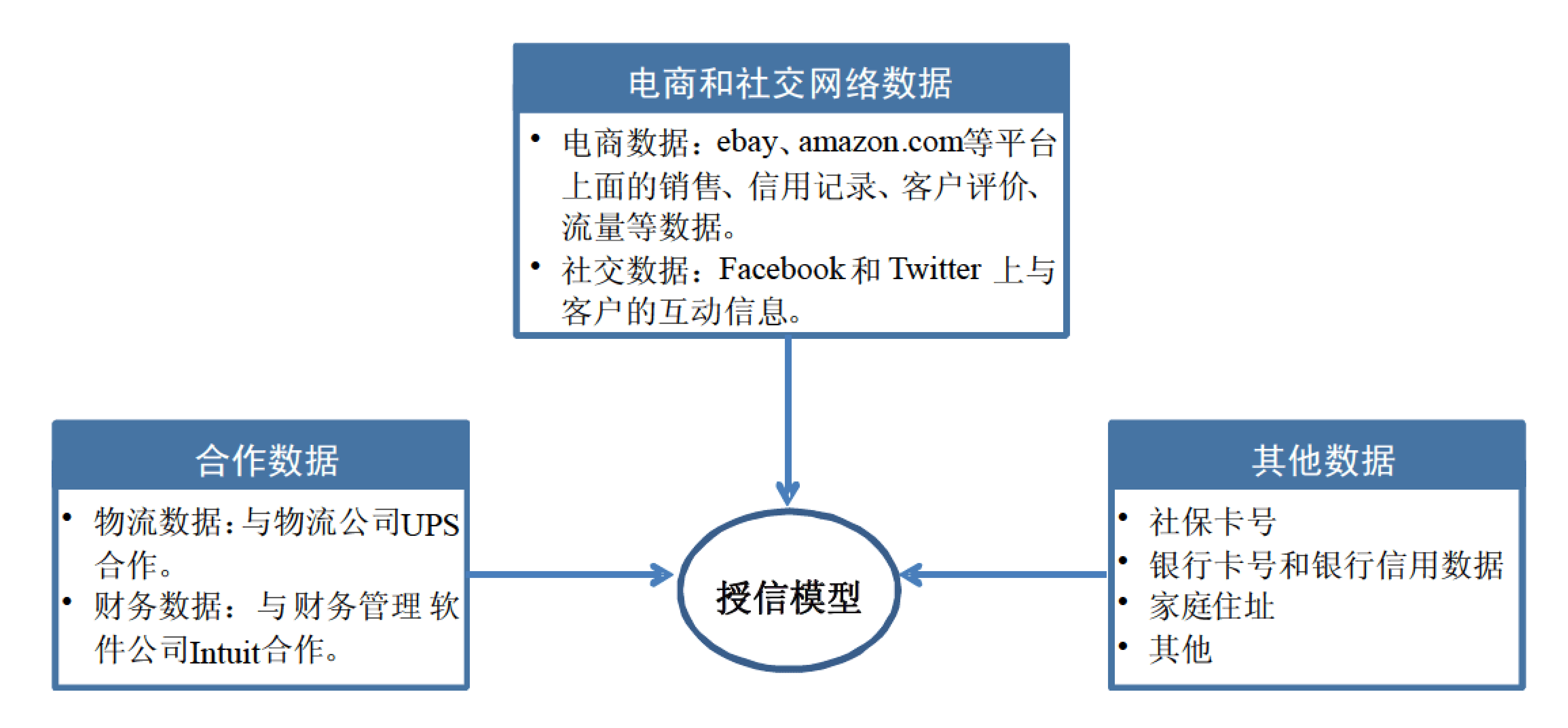


图 5-8　美国 Kabbage 授信模型的主要数据构成

美国 Kabbage 公司商业模式充分体现出了信用数据化和数据资产化的发展趋势。未来，随着数据规模的快速膨胀，数据资源的获取将不再成为企业竞争的关键优势，数据的分析利用能力成为竞争的焦点，当然这种状况需要数据共享和数据所有权归于用户两个条件的成熟。

（3）金融机构体系重新构建，提升结构效率

传统资金融通方式在促进资源配置和经济增长的同时，也产生了巨大的交易成本。目前中国资金融通主要是通过交易中介和服务中介两类中介机构，交易中介主要包括银行、证券公司等机构，交易服务主要包括会计事务所、律师事务所、投资咨询公司等机构。交易中介提供两种融资方式：一是通过银行的间接融资方式，这

也是中国当前主要的资金融通方式；二是通过证券公司进行股票或债券的直接融资方式，如图 5-9 所示。这两种资金融通方式对于促进经济增长和资源配置起到了非常重要的作用，但同时也产生了巨大的交易成本，直接体现为银行等金融机构的利润，据中国银监会的统计显示，2012 年中国商业银行的净利润近 1.24 万亿元，较 2011 年增长了近 18.96%，远高于 2012 年 GDP 增长速度 7.8%。

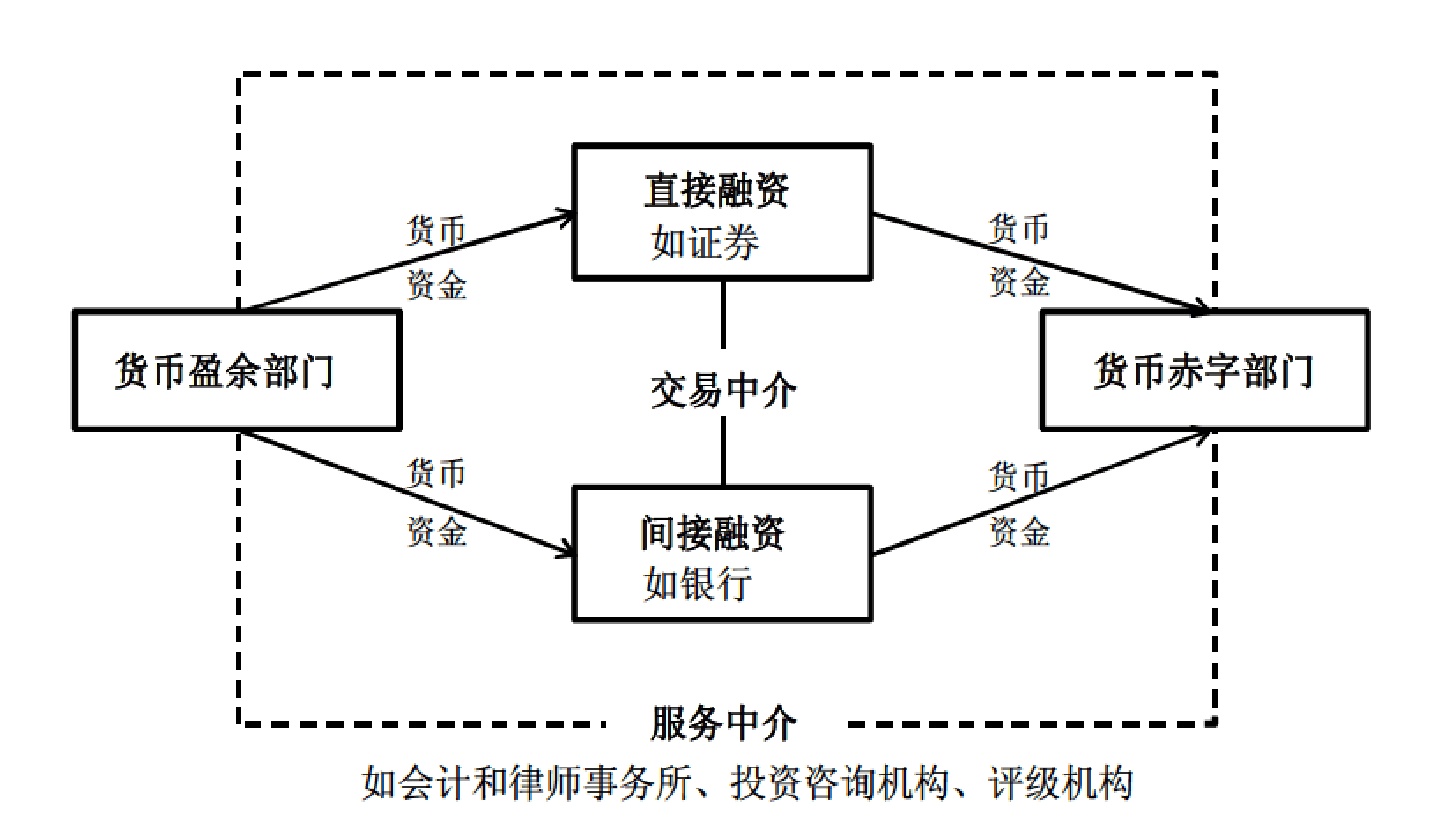


图 5-9　中国传统金融市场资金融通方式

银行等金融中介存在有两个主要前提：一是交易费用的存在，金融交易是跨地域跨时间的，不确定性更大，旨在降低风险和不确定性的交易费用更高，金融中介通过专有技术可以实现规模经济；二是信息不对称的存在，导致逆向选择和道德风险，金融中介通过信息生产加工和账户监控等方式可以缓解上面的逆向选择和道德风险两个问题。

搜索引擎、社交网络、物联网、移动互联网、云计算、大数据等新兴信息技术改变了传统的信息产生、传播、加工利用的方式，打破了信息不对称，降低了信息获取和加工成本，这将加速交易中介的脱媒化进程。

未来的金融模式将是资金供求双方实现自由匹配，且是双向互动社交化，如图 5-10 所示。但金融业不仅存在信息不对称，同时也存在知识不对称，金融产品具有风险性特征，因而个性化的解决方案咨询仍有市场。不过 IT 可以将人类知识结构化，且随着机器学习、IT 智能的发展，服务中介的部分功能也会逐渐被 IT 智能支持所取代。目前新兴的 P2P 模式和众筹模式已经显示出金融业的这种发展趋势。

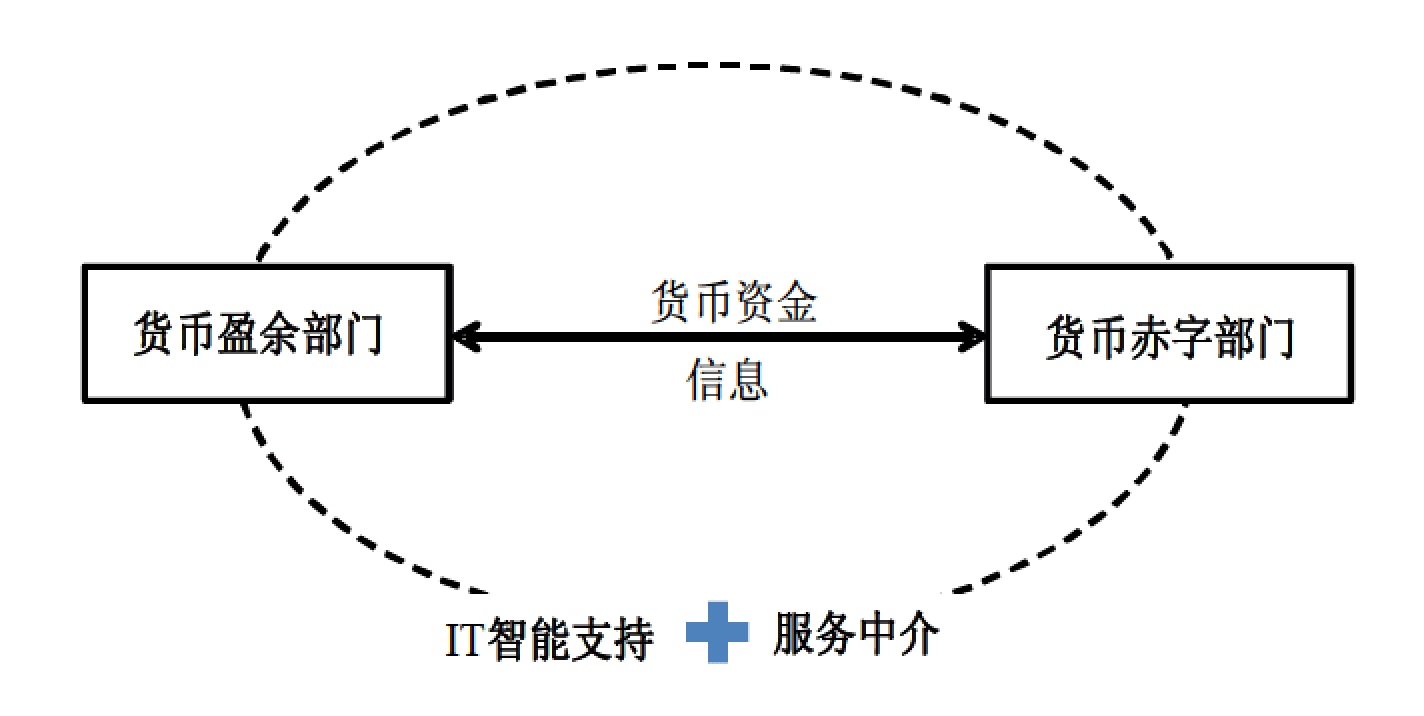


图 5-10 未来的金融运作模式

P2P 模式冲击传统银行中介模式。P2P 网络借贷与 eBay 的运作模式比较类似，即为贷款人和借款人搭建了一个展示、交易的网上平台，拟借款人需要填写贷款金额、用途、期限、信用记录以及个人信息等资料，网站会对拟借款人的资料进行初步审核并给出量化的信用评分和风险等级，拟贷款人可以据此设计投资方案。在国外 P2P 模式发展态势良好，以美国 Lending Club 为例，截至 2013 年 1 月初，贷款总额已经超过 12 亿美元，创造利息收入突破 1 亿美元。P2P 引入中国虽然相对较晚，但发展却极其迅猛，宜信、拍拍贷、红岭创投等一批新兴 P2P 网络借贷公司应运而生，同时，中国平安集团（陆金所）、国家开发银行（开鑫贷）等传统巨头也纷纷参与角逐 P2P 市场。

众筹模式冲击传统证券中介模式。众筹模式是通过网络平台面向公众筹资，该模式的兴起源于美国的 Kickstarter，该网站是一个创意方案的众筹平台，人们可以通过该平台向公众募集小额资金，用以实现自己的梦想，截至 2013 年 4 月，Kickstarter 已经帮助人们获得了 5.71 亿美元融资。目前可以初步将众筹模式分成四类：一是生活众筹模式，如 Crowdtilt，筹资主要是用于满足生活小梦想，如举办一个集体活动；二是股权众筹模式，如 FundersClub，筹资主要是用于帮助人们创办一家企业；三是产品预售模式；四是公益模式。2012 年 4 月，奥巴马总统签署了《促进初创企业融资法案》，对众筹模式进行了定义和规范，为众筹模式的发展提供了法律支撑。然而，在中国，虽然也出现了点名时间等众筹模式网站，但总体上还尚处于萌芽状态。

重构金融格局

（1）中国金融业面临三层竞争

新兴信息技术和国家大资管政策促使中国金融业出现三层竞争：一是金融业的

潜在进入者与传统各类金融机构之间的竞争；二是银行、保险、证券和基金等传统金融机构之间的直接竞争开始加剧；三是全国大型金融机构与区域中小型金融机构之间的正面竞争日趋激烈，如图 5-11 所示。

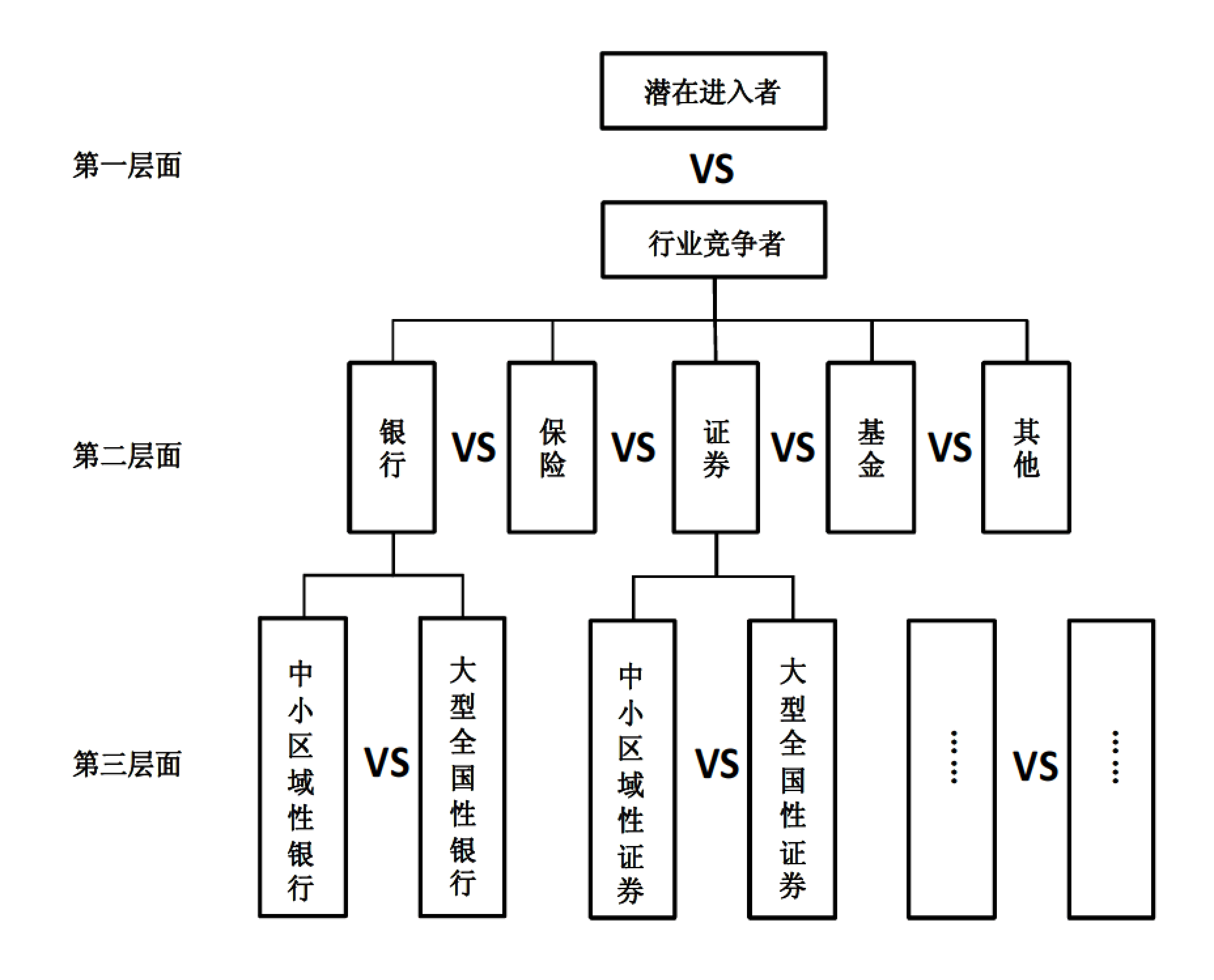


图 5-11　中国金融业的三个层次竞争

大量金融业的潜在进入者凭借互联网和大数据涉足金融业，打破了原有的竞争格局体系。金融业的潜在进入者可以分为两类：一类是跨界企业，主要是以阿里巴巴、京东商城、谷歌等互联网企业为代表，新兴技术推动产业边界日益模糊化，跨界竞争日趋常态，这类企业依托在各自领域的多年积累，掌握了大量的用户数据，并借此进入金融业满足用户的金融需求，推动生态体系的发展；另一类是基于互联网的初创企业，具体包括支付宝、财付通等第三方支付企业，宜信、Lending Club 等 P2P 网络借贷企业，Kabbage 小额网络信贷企业等，如图 5-12 所示。

大资管政策推动金融机构的混业竞争。以前，银行、保险、证券、基金等金融机构是分业经营，在各自领域攫取利润，而今混业经营已经成为发展趋势，传统金融机构之间的竞争也将日趋激烈。银行在中国金融体系中处于强势地位，证券、基金等金融机构若想在混业角逐过程中取得胜利，互联网将成为其关键利器。

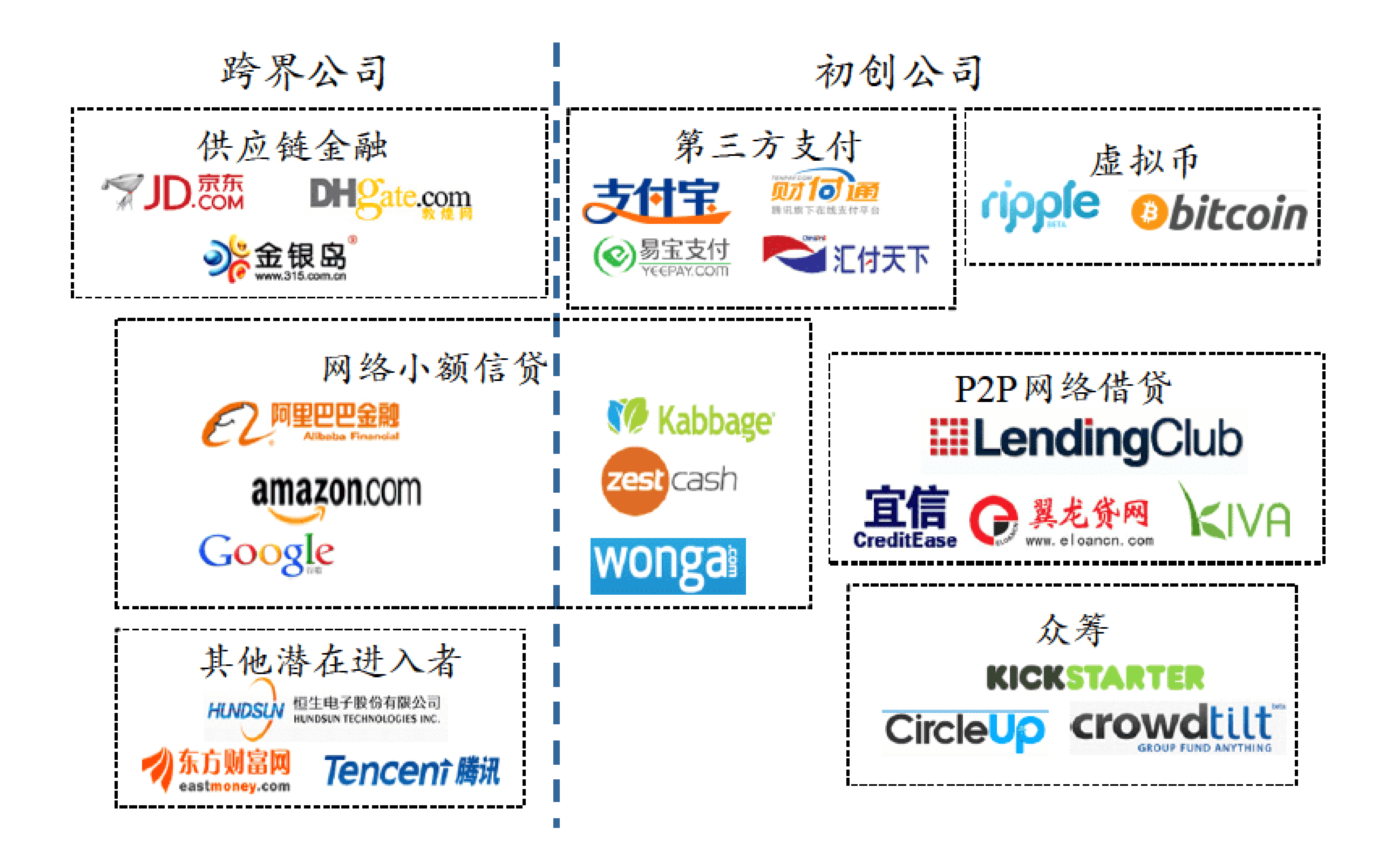


图 5-12　金融业潜在进入者构成

互联网和大数据打破了信息不对称和物理区域壁垒，使得中小型、区域型金融机构与大型、全国型金融机构站在同一层次竞争，迫使中小机构转型开展差异化竞争，否则将难逃被淘汰的结局。以证券公司为例，之前很多区域证券公司凭借区域优势收取较高的经纪费率，但在 2013 年 3 月，中国证券登记结算公司推出《证券账户非现场开户实施暂行办法》，允许用户通过网络进行开户，这将对区域证券公司带来较大的冲击。

（2）大平台+众多小而美的产业格局是未来方向

金融业的三个层次竞争将推动产业格局重构，大平台+众多小而美的格局将成为未来发展趋势。在大数据时代和混业竞争的背景下，实力强的大型企业将大肆扩张，由于金融业信息密集型的特点，大平台将凸显赢者通吃的态势，尤其在标准化产品和低净值客户领域将更加凸显其规模优势和成本优势；与此同时，其他实力较弱的企业被迫寻求差异化竞争的道路，改造和转型线下传统营业厅，通过线上线下深度融合的方式重点针对高净值客户提供非标准化产品和服务，否则将难逃被淘汰的命运，由于金融业知识密集型的特点以及多层次的金融需求的存在，在一些细分领域市场仍将有很大的生存空间，如图 5-13 所示。

未来金融业的参与者中将既包括传统金融机构（一批被淘汰，一批转型后生存），又包括互联网企业跨界者和初创企业，至于哪方力量占据主导地位目前尚不能给出明确定论，而且这种判断意义也不大，因为未来两方力量终将殊途同归，最终存活下来的企业必然兼具金融和互联网两方面的基因与能力。

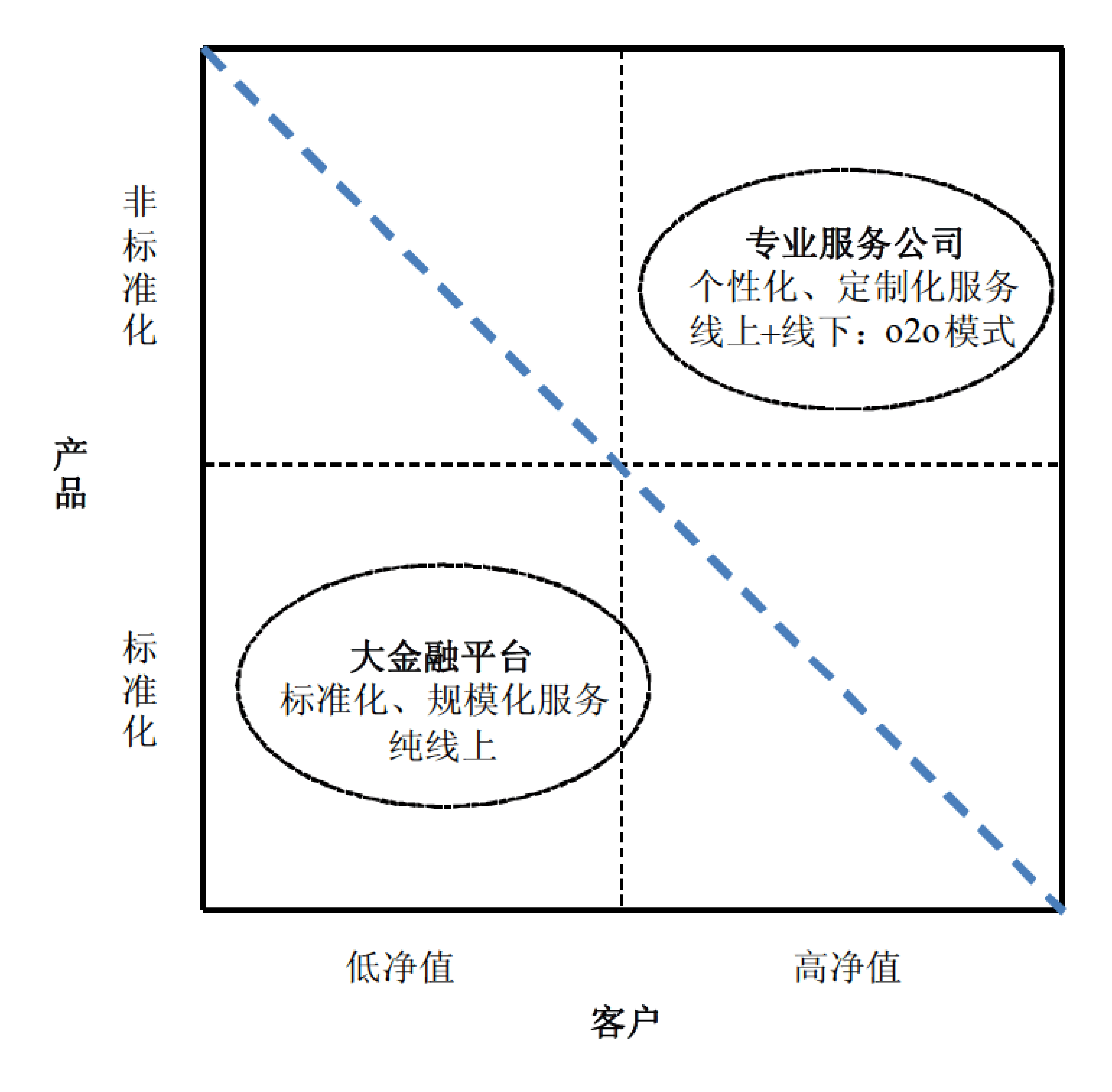


图 5-13　未来金融业的产业格局

第四节　中国互联网金融将引领全球

中国经过三十年的改革开放发展，取得了令世人瞩目的成绩。如今，中国正处于一个新的转折点，大数据所引领的第三次工业革命很可能会在中国爆发，由于中国的金融业基础薄弱、网民基数巨大，中国 IT 技术与金融业融合的程度和速度都会超过美国，推动中国经济新一轮的高速增长，由此 C2C（Copy to China）的模式也已经过时。

一方面，中国利用 30 年时间几乎完成了西方发达国家百年的发展过程，虽然取得了惊人的业绩，但总体上产业基础还不是非常健全，同时，中国金融业由于长期政策垄断，使得产业效率比较低，市场竞争力有限，这为互联网和大数据提供了广阔的可颠覆空间。

另一方面，中国拥有广大的网民基础，有利于快速形成规模，且大部分网民已经养成了网上娱乐、消费的习惯和行为。中国无论是手机用户还是固定宽带网民用户都已经远远超过美国，且目前的增长速度也明显高于美国，这为中国互联网行业发展奠定了坚实的用户规模基础，以腾讯推出的微信为例，微信仅仅发布两年，用

户规模就已经达到 4 亿用户，掌握了巨大的流量和数据后，商业变现能力大幅提升。

事实上，这种趋势在零售业已经得到了印证。2012 年中国电子商务的渗透率是 6.3%，已经超过美国电子商务的渗透率 5.17%，且中国电子商务的渗透率仍然保持高速增长，由此可见，互联网和大数据对中国零售业的颠覆程度将远远超过美国，如图 5-14 所示。通过对比中美两国的电商巨头亦能发现这一趋势，2012 年阿里巴巴集团旗下的天猫和淘宝两大平台的网络零售交易额已经突破万亿元，在 2012 年的第四季度，阿里巴巴的商品成交总额超过亚马逊和 eBay 的成交总额之和，如图 5-15 所示。

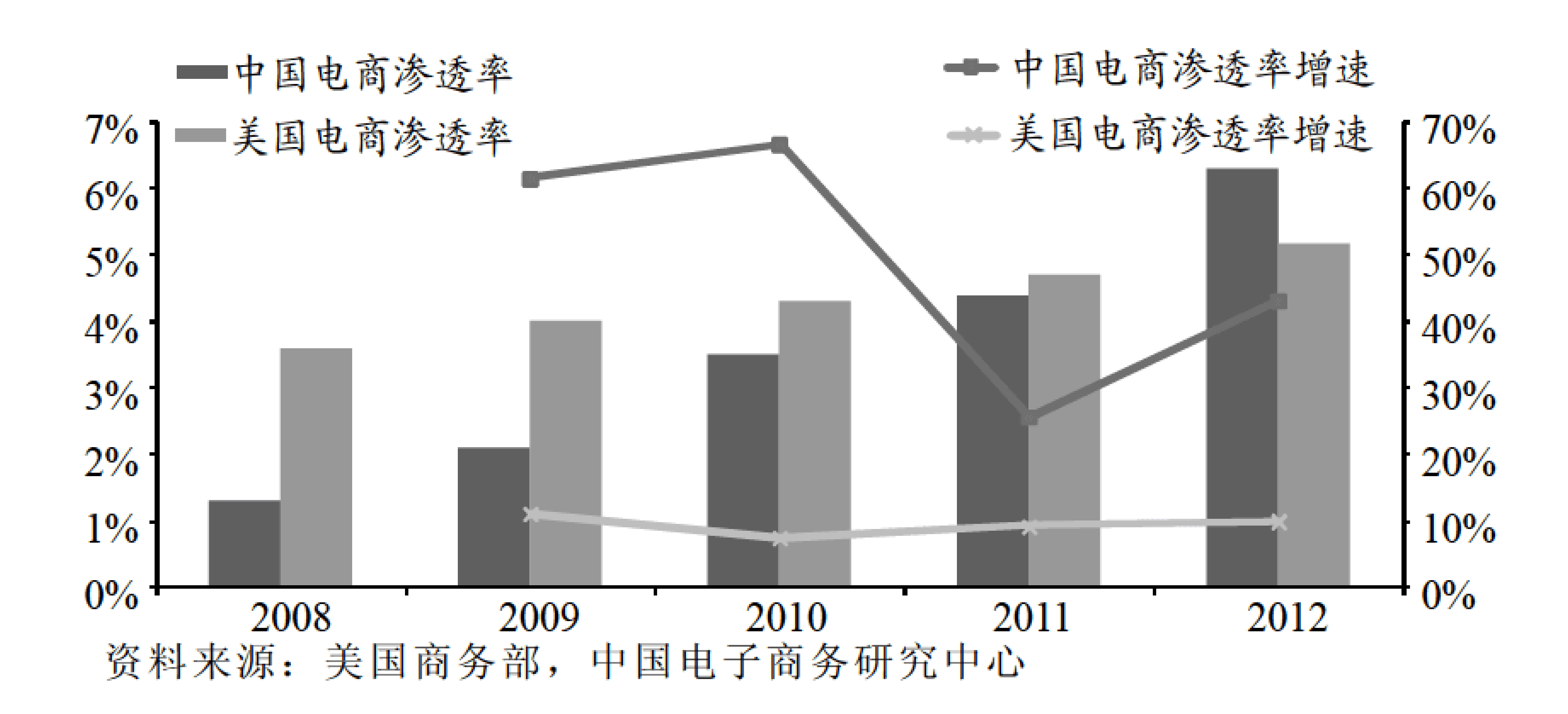


图 5-14　中美电商渗透率对比

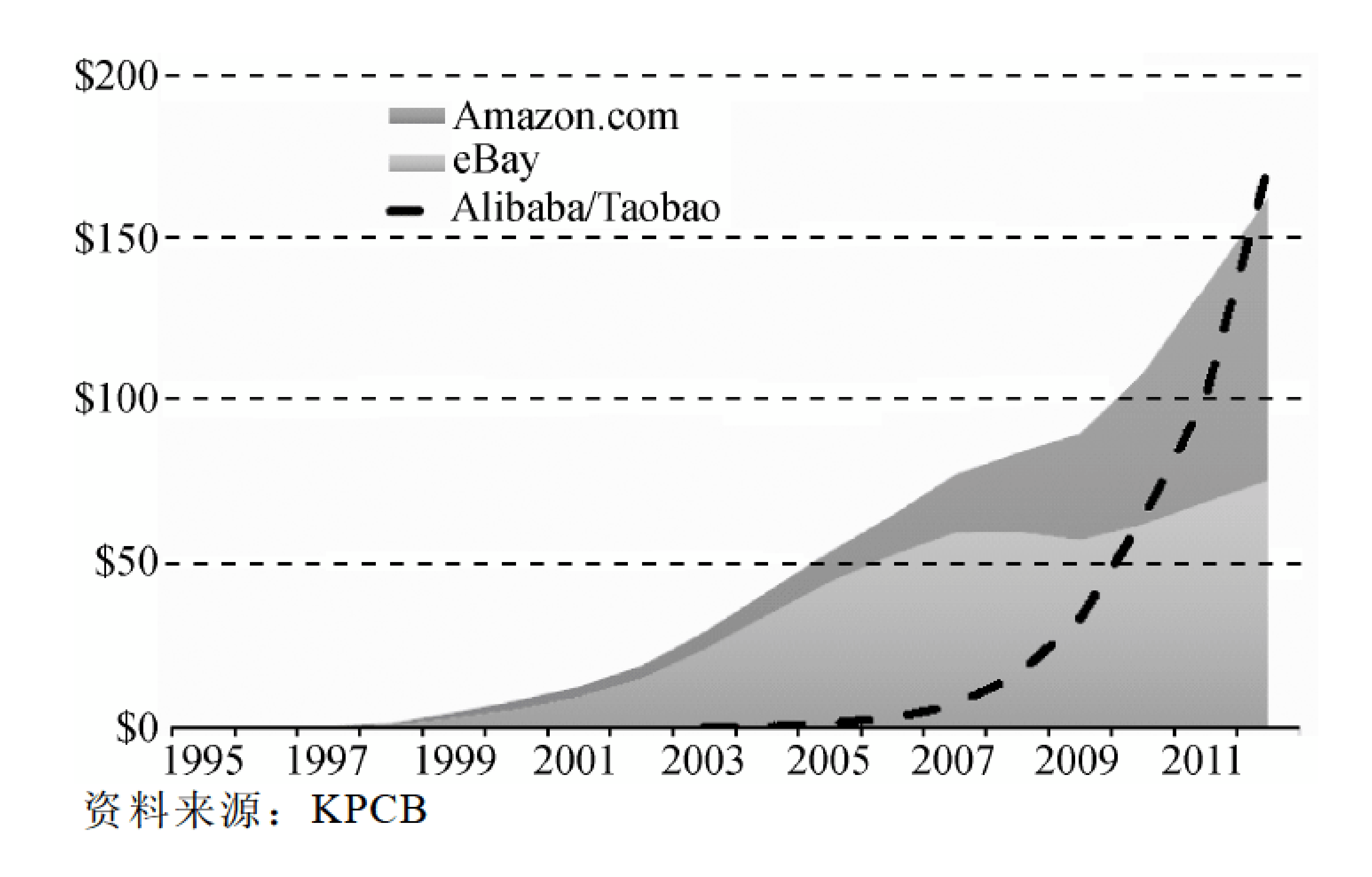


图 5-15　1995—2012 年亚马逊、eBay、阿里巴巴/淘宝商品成交总额对比

相信不仅仅是零售业和金融业，未来医疗、教育等众多产业都将受到互联网和大数据的影响，而在这一轮变革中，给予了中国弯道超车的机会，中国极有可能成为世界的领导者。

导读：

1. 借助大数据技术，公司可以比以往任何时候都更加了解消费者。那些拥有大量消费者，并能洞悉消费者行为的公司，开始掌控产业链。另一方面，产业链上游公司则积极向下游渗透，争夺话语权，宏观层面体现为产业内部垂直整合的趋势不断加强。
2. 在产业垂直整合的大趋势下，无论是消费市场还是企业服务市场，产品层面的软硬一体化成为信息产业垂直整合的微观表征。我国信息科技的发展初期，一度重硬轻软，政府为扶持软件企业发展，不得不矫枉过正。走过一个轮回后，软硬一体化，或许是从事信息服务的企业做大做强的契机。
3. 体验为王、大道至简，成为判断一体化产品能否获得消费者青睐的关键因素。

第六章

大数据加剧产业的垂直整合趋势

越靠近最终消费者或用户，在产业链上就拥有越来越大的发言权。

——笔者

行业内的分分合合，大致遵循相似的发展规律。上游企业的产品，是下游企业的原材料，一环扣一环，最终交付到消费者手中。行业的每个环节都有数家公司在竞争，同环节竞争非常惨烈，常常是有你无我的零和游戏。在同一个产业环节中存在的收购整合，一般称为“横向整合”。横向整合的结果，会形成在某产业环节的垄断，或者几家均势的竞争格局。这是企业做“大”的过程。

在某环节占据优势地位的公司，往往开始沿产业链上下游展开收购整合。不同行业，不同的历史时期，战略节点亦不同。有些公司向下游扩展，有些公司则溯流而上，向上游扩展。这种沿产业链上下游展开的收购整合，一般称为“垂直整合”。垂直整合可以看作是公司做“强”的过程。

“大”和“强”的概念都是相对的。非“大”无以至“强”，非“强”无以至“大”，“大”且“强”者方能生存。公司在某个产业环节做大了，但还是可能受到上游供应商，尤其是核心部件、垄断性资源、关键技术的限制。虽然是大公司，面对上游强势的资源、技术、产品的公司，仍表现为弱势的一方。长期以来，我国的个人计算机制造商就是这样的行业地位，一直受上游操作系统厂商——微软和芯片厂商——英特尔的限制。同样，如果产业链下游存在大型的渠道商，上游的制造商需要依赖渠道商把商品交付到最终消费者手中，则制造商在和强势渠道商的博弈中也会处在弱势的地位。

产业垂直整合是公司由大而强的必经之路。号称“共和国长子”的中粮公司就是全产业链模式的代表。中粮把小麦、玉米、油脂油料、稻米、大麦、糖、番茄、肉食等产业链有机组织起来，实现了整个粮油食品链条从种植到食品营销的畅通无阻。苹果公司也一直牢牢把握从芯片设计到苹果零售店等所有产业链的关键环节，为用户提供完美的购买、服务和使用体验。

产业内的垂直整合趋势，随着技术的发展、各环节博弈能力的此消彼长，逐渐呈现下游的公司挟消费者这一“天子”以令上游诸侯的局面。产业链上的战略节点，逐渐向消费者端迁移，形成以消费者为中心的产业格局。

第一节　形成以消费者为中心的产业格局

提要：

1.“新一代基于互联网 DNA 企业的核心能力在于利用新模式和新技术更加贴近消费者，深刻理解需求，高效分析信息并做出预判，所有传统的产品公司都只能沦为这种新型用户平台级公司的附庸，其衰落不是管理能扭转的。互联网的魅力就是‘the power of low end’”——出井伸之，索尼公司前董事长

2. 中国企业界，除了百度、腾讯、阿里巴巴这些互联网公司外，小米公司是践行以消费者为中心的代表。小米和小米粉丝之间，制造商和消费者的天然鸿沟在消退。一些铁杆粉丝随着小米的壮大，慢慢成了小米的员工。即便不是其员工，也可以通过社区介入到小米手机的设计和测试环节中去。这两大群体，通过精心运营的网络媒介，形成互相促进的两大力量。

社交网络、大数据等新技术的应用，极大地放大了消费者的博弈能力。制造能力的进步，使得上游厂商的产品同质化程度加剧。它们不得不更贴近消费者，要更主动倾听消费者的需求，才能在竞争中胜出。这些因素叠加在一起，客观上推动了以消费者为中心的产业格局变迁。索尼这家以供应链管理见长的制造型公司，比起做网上书店起步的在线零售帝国——亚马逊公司，对消费者需求把握的时效性和精准性，实在是云泥之别。

德鲁克[①]的经典问题——你的客户是谁?

德鲁克的经典著作《管理：使命、责任、实务》奠定了 21 世纪管理学的基础。

① 彼得·费迪南德·德鲁克（Peter Ferdinand Drucker，1909－2005）是一位奥地利出生的作家、管理顾问以及大学教授。他催生了管理这个学科，被誉为“现代管理学之父”。

他老人家经常问企业高管的第一个问题就是“你的客户是谁？”，这个问题看起来非常简单，但是核心在于你真的了解你的客户吗？

这是微博上流传的一个段子：“如果你没有跟客户吃过饭、K过TV，不要告诉我你的客户将要下单；如果你的客户兜里没有塞了你给他的XX，你不要告诉我你的客户会持续下单……”。郭台铭的语录：“如果你不知道客户的内裤颜色，不要告诉我他会把业务给我们做……”。我们无意对此问题做商业伦理上的评价，只看其中传递出来的一个强烈主旨：你这样做了，我也许可以认为你了解你的客户，包括他的业务需求和个人爱好。

当面对10亿消费者的时候，我们如何去了解每个人的兴趣和爱好呢？

亚马逊的实践

杰夫·贝索斯[①]开会时常留出一把空椅子，提示与会者未在场的消费者才是最重要的人。贝索斯是世界上第一家网络书店亚马逊（Amazon）的掌门人，其11年财报显示，亚马逊营收480亿美元，净利润6.3亿美元，市值接近千亿美元，成为互联网界新的精神领袖。

在亚马逊，推出新的产品或者服务，不需要经过冗长的调研、分析、讨论等环节，而是尽可能快地推出产品。短短两周内，消费者就会在公司网站留下访问、评论、购买、推荐等各种数据。接下来就是大数据技术出场，分析这些海量的数据，评估产品是否令人满意，预判消费者是否会为类似产品慷慨解囊，从而决定这款产品或者服务是否应该继续推向市场，或者应该取消，启动另外的尝试。

这种决策流程的变化，真正地把消费者置于整个企业决策的中心地位，这也是为什么贝索斯开会时会空出那把椅子的原因。因为他深刻的理解到，在互联网时代，消费者开始具备了颠覆的能力。

---

① 杰夫·贝索斯（Jeff Bezos），创办了全球最大的电子商务公司之一，亚马逊。他是全球电子商务的第一象征。1999年当选《时代》周刊年度人物。

把消费者话题放到社会时代变迁的背景来研究，则更容易理解。当今世界正处于从工业化向信息化过渡的时代，美国要快一点，中国紧随其后。工业化主导的特征是大生产、大物流、大品牌、大零售，通用汽车、UPS、宝洁、沃尔玛是大工业时代的代表，如图 6-1 所示。而未来是消费者主导的信息社会，其以消费者驱动、个性化生产、网络化协作为特征。过去企业的发展只要专注内部的生产、管理、供应链等问题就够了，消费者只是被动的接受。未来企业的内涵扩展，边界消弭，消费者将成为企业重要的一份子。

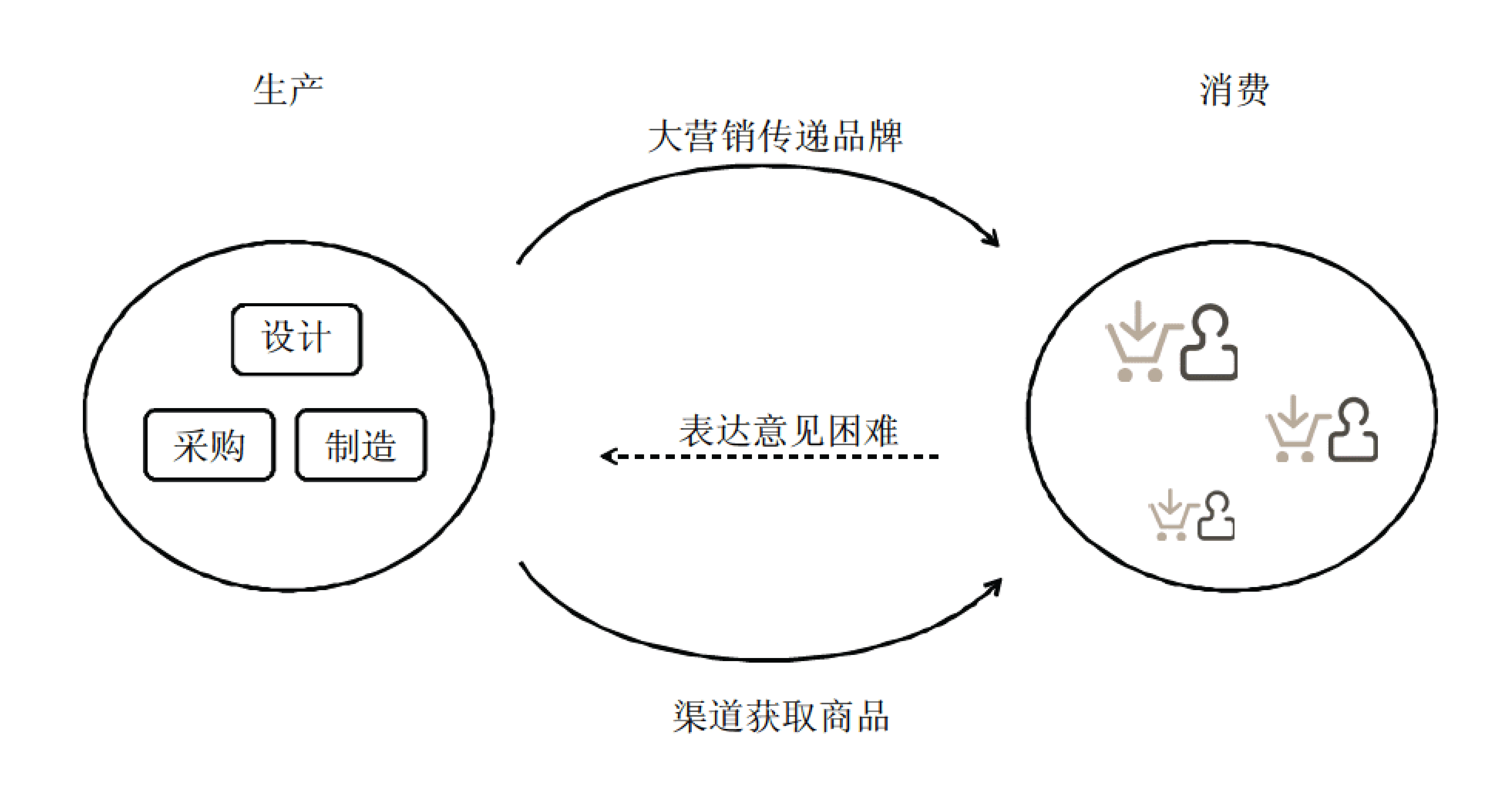


图 6-1 工业时代以生产为中心

譬如，福特汽车公司早期生产的 T 型车[①]，福特曾笑言“顾客可以随意选择他喜欢的颜色，只要是黑色”。这是典型的以生产为中心的商业模式，以当时生产线的技术能力、工艺水平，如果增加不同的颜色，会显著影响生产效率和制造成本。但现在制造业水平突飞猛进，具备个性化生产的能力。最新的印刷技术，印制 1 万张相同的图像和印制 1 万张不同的图像成本近乎相同。这和福特 T 型车时代，呈现完全

① 福特 T 型车是美国亨利·福特创办的福特汽车公司于 1908 年至 1927 年推出的一款汽车产品。第一辆成品 T 型车诞生于 1908 年 9 月 27 日，位于密歇根州底特律市的皮科特（Piquette）厂。它的面世使 1908 年成为工业史上具有重要意义的一年：T 型车以其低廉的价格使汽车作为一种实用工具走入了寻常百姓之家，美国亦自此成为了“车轮上的国度”。

不同的商业图景。当下，必须洞察消费者的喜好，而且是每一个消费者的喜好，才有可能提供个性化的产品。

大数据技术的发展，开启了这扇洞悉消费者心理的方便之门，如图 6-2 所示。曾经有服装企业想调查其顾客的购买意愿，看哪件衣服顾客拿起来了，哪件试过了，又要安摄像头，又要选样本，没有小一亿下不来，要想省钱减少样本量，可能又会面临统计结果失灵的风险。但在互联网上做同样的事情，成本近乎于“零”。因为消费者在网页上停留的时间、点击衣物图片、放到购物车等行为无一不清晰的记录在服务器上。分析这些数据的唯一挑战，就是迅速的在海量数据中形成有助于决策的信息。而这恰恰是大数据技术发挥价值的领域之一。

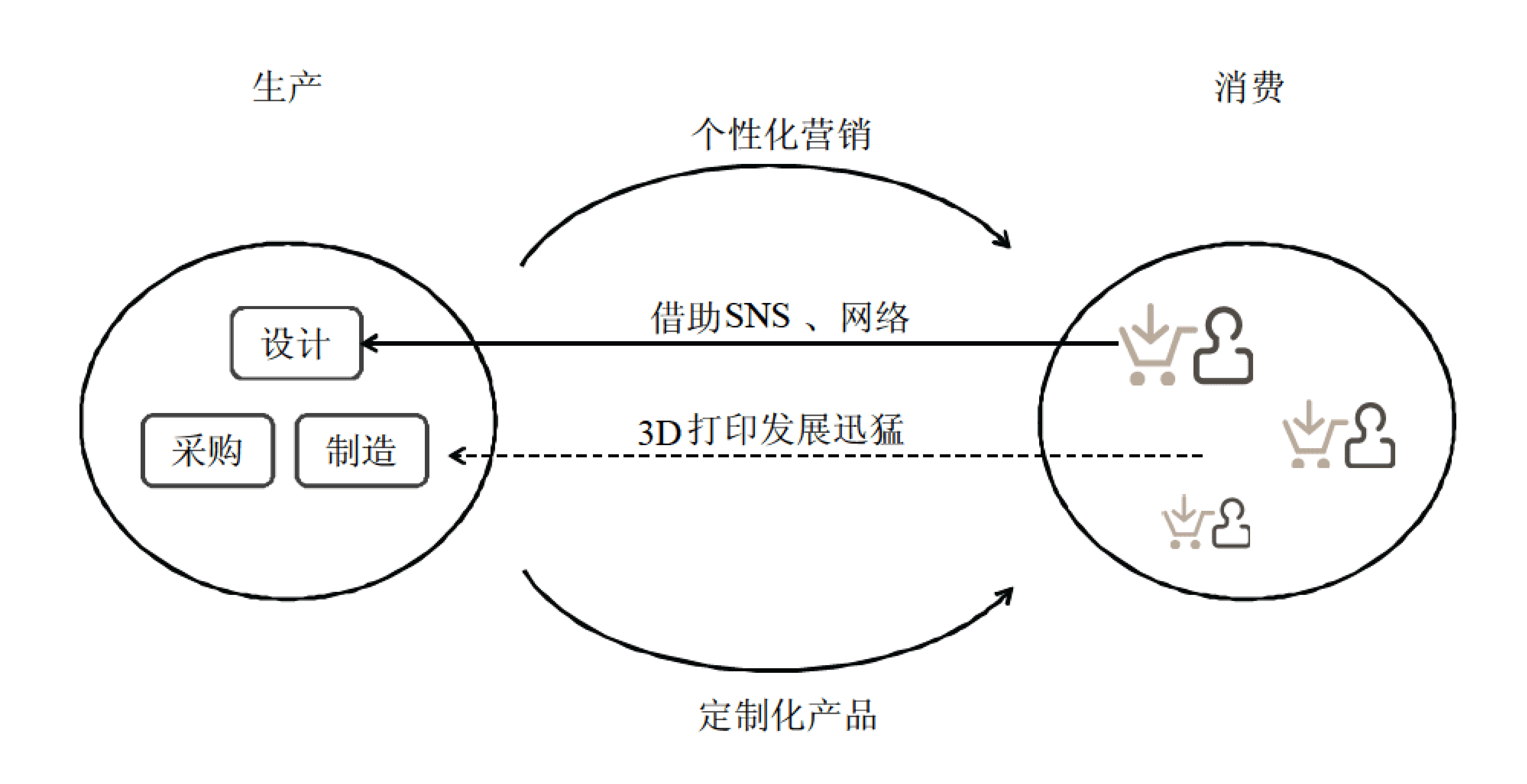


图 6-2　以消费者为中心，引发产业融合、企业变革

洞悉消费者的心理，准确、快速响应消费者需求，甚至是潜在需求，是当代制造业向个性生产转型的关键。拥有海量数据处理、分析的技术，将是这些企业的千里眼和顺风耳。目前，对于国内大部分企业来说，大数据的商业价值正处于启蒙阶段。令人欣慰的是，总有一些具备领袖气质的公司，走在技术应用的前沿，也总有一些企业家在不断的学习和超越。

回到这一小节的标题上来，对比一下亚马逊和索尼。亚马逊利用庞大的电子商

务网站，大规模收集消费者在购买商品过程中留下的点击、评论、购买等数据，精确预测消费者的兴趣点。并以数据为核心，开始涉足平板电脑，低价提供 kindle fire 系列产品，成为目前市场上唯一一款可以匹敌苹果 iPad 的利器。索尼近些年推出的产品，明显远离消费者，记忆中已经没有任何产品可以再现 Walkman 随身听的盛况。在大数据时代，索尼明显落伍了，它错过了利用互联网搜集用户数据，更加贴近消费者的历史机遇。

恒安国际的核心竞争能力向消费端迁移

如果跳出信息产业，把所有的产业都抽象地描述为“生产”、“消费”两大环节的话，大部分产业的主导力量，都在向“消费”端迁移。零售业一直是以消费者为中心的产业，大家非常容易理解，现在就简单剖析一个消费品制造业的例子，一探究竟。

恒安国际是一家生产纸制品的公司，旗下产品包括纸巾、卫生巾、纸尿裤等产品，占中国大陆市场第一名，目前销售收入已经超过 170 亿港元。该公司于 1998 年在香港上市，是恒生指数的成份股。恒安国际的老对手是宝洁公司，在纸巾这个领域，宝洁最终不敌恒安，彻底退出中国市场。恒安国际的能力可见一斑。

在和恒安国际的董事会交流大数据的影响时，许连捷总裁分享了他们是如何在产业竞争要素变迁的大背景下，把宝洁一步一步赶出中国市场的过程。略去其中精彩的商战故事不谈，来看看不同的历史阶段，哪些是事关制造业生死的核心命题，恒安又是如何化解的。

最初恒安生产的纸巾，质量比较差。拿来擦脸的时候，胡子茬会刮下来纸巾上的碎屑，没准个把黑胡子就变成了白胡子，产品质量和宝洁公司根本不是一个层次。这个阶段，谈精准营销、供应链都是不着边际的。当务之急，是要把产品的质量提上去。恒安开始大举投入全面引进德国的生产设备，改造生产线，培训工人，全力以赴地改善产品质量。

很快，恒安和宝洁的产品在质量上已经难分伯仲。这个时候，拥有经销商数量的多少，成为决定两家市场份额的关键因素。大力地发展优质经销商、区域总代经等合作伙伴，是公司跨过质量门槛后的第二大考验。

真正和宝洁拉开距离的是恒安对销售终端的控制。恒安拥有 3 万人的销售代表，活跃在大大小小的超市等销售终端，消费者在哪里扎堆，哪里就有恒安的销售代表。管理庞大的本土化的销售代表，不是宝洁的强项。竞争至此，宝洁选择退出中国大陆的纸巾市场。

回顾恒安这段历程，公司事实上在一步一步向消费者靠近。在制造业，同样是“越是靠近最终消费者，在产业链上就拥有越来越大的发言权”。这样的故事在零售业不断上演。前几年以国美、苏宁为代表的强势家电零售商，不断侵蚀压榨家电制造商的利润。家电制造商有苦难言，却又不得不低头。现在以京东商城、天猫为代表的电子商务公司，同样扼住了制造商的咽喉。原因无他，因为这些公司更懂得消费者、更贴近消费者。

恒安国际已经有了 3 万人的销售代表，他还有什么手段更了解消费者呢？传统的物理的方法，可以说已经被恒安发展到了极致。但是和电子商务公司比起来，就像刀耕火种时代的民兵，遇到全副武装的空军一样。恒安的老板意识到大数据的巨大价值，最后总结了一段非常有代表性的话：“我们必须要搜集消费者的购买数据、关注数据，来改善产品的设计和销售，降低库存，优化采购。一句话，就是要把大数据融入到企业的经营中去。”

小米的粉丝文化

“业界对小米的看法经历了三个阶段，起初是看不起，后来是看不懂，到现在是赶不上。”这是小米公司一位高管接受记者采访时说的一段话。小米手机的发展的确超乎所有人的意料。小米是完全围绕消费者来经营的，更确切地说，小米围绕他的粉丝们来经营。这种经营思想的转变对许多公司都有巨大的参考价值。这节内容是

根据作者今年 6 月份的一篇博文改编，当时许多人都以黑小米手机为乐。

最令人关注的是，他是如何聚拢 300 万的粉丝团的？粉丝在小米的发展过程中，有什么作用？

根据官方资料，小米最早推出的产品就是 MIUI 手机操作系统，根据 Goolge 安卓系统定制而来。MIUI 操作系统受到一些玩家的喜爱，一些人买的安卓手机就直接刷机成 MIUI。同期推出的小米论坛，聚拢了一些铁杆的 MIUI 用户，他们中的有些人，从用户到粉丝，到最终加入小米工作，成为论坛版主和运营人员。

这些铁杆粉丝可能天生喜欢折腾操作系统，同时他们在其朋友圈里也是公认的手机专家，别人遇到智能手机的问题，往往也会找他们解决。因此，这些铁杆粉丝其实具备了影响他人的能力，成为小圈子里面的手机领域的意见领袖。

MIUI 受到铁杆粉丝的追捧，也是和小米的快速升级策略紧密相关。想想看，当你抱怨某个功能不完善或者出现错误的时候，小米团队立即作出反应，在下一个版本中改正了这个缺陷，甚至在论坛中大力褒扬提出问题的粉丝。每个提出问题粉丝，都有了一种近乎神圣的参与感。也许他们没有亲手参与开发过程，但是充当了需求方、系统检测方、甚至是部分功能设计者的角色。如此一来，小米手机不仅仅是小米团队的手机，而且是小米粉丝们的手机。铁杆粉丝的深度参与，使得他们对小米手机有天然的亲切感。尽管 MIUI 问题不断，但是哪部手机会没有问题呢？况且乐趣就在于亲身解决问题的过程中。

白酒营销中，有人提炼出盘中盘的营销思想。说白了就是通过公关某地显要阶层，以此阶层辐射带动其他阶层的营销手段。小米的营销带有明显的盘中盘特征。

譬如，在 798 这种充满艺术气息的场所举行新机的发布会，捧场的近千人都是小米的粉丝。营销商、合作伙伴反而成了少数人。这些粉丝直接就会变成新机的用户。据说同期还有另外一家手机厂商，在人民大会堂举行新机发布会，请到工信部、运营商领导、明星助阵，声势浩大，但现在几乎没有人知道这家公司。上网倒是可以搜到这些新闻，但单向传播在微博时代已经没有多大的意义了。

反观小米对粉丝的运营和发掘，可圈可点。这些粉丝有松散的组织。小米通过

社区解决和铁杆粉丝深度沟通的问题。社区中不但有各种技术贴，更是手机推广和销售的主渠道。小米粉丝和小米用户是高度重叠的。小米有大约 20 多人的团队专门负责运营社区。此外微博是互动的另一个高效平台。

小米手机官方微博的粉丝大约在 300 多万，而其手机累计销量也是 300 多万，这两个数字可能是巧合，但也能反映其粉丝和用户重叠的现象。小米也非常注重微博运营，大约有 20 人左右负责小米的官方微博。

小米之家担负了线下品牌传递的重任。在小米之家，用户可以体验新功能，解决手机故障、维修等问题，类似苹果商店。

小米和小米粉丝之间，制造商和消费者的天然鸿沟在消退。一些铁杆粉丝随着小米的壮大，慢慢成了小米的员工。即便不是其员工，也可以通过社区介入到小米手机的设计和测试环节中去。这两大群体，通过精心运营的网络媒介，形成互相促进的两大力量。

这是一种新型的制造商和消费者的关系。具备了后工业时代“以消费者为中心、定制化生产、网络化协作”的雏形。小米的组织结构中，客服部门至关重要，微博、社区、小米之家，都属于广义的客服部门。

这种模式值得许多生产、制造型的企业效仿，不能仅仅倾听用户的声音，要让用户介入到你的设计、制造、营销、反馈环节中去。现在大多数的企业官方微博，仅仅是个传声筒，甚至沦为摆设，这样的公司的投资价值明显要小于善于和粉丝们打交道的公司。

第二节　信息产业的垂直整合趋势

提要：

1. 具体到信息产业内部，上游同质化的表征是开源软件的兴盛。所有重要的商业软件都有对应的开源版本，所有大型的互联网公司，几乎完全依赖开源软件运营。

2．企业信息化市场，产业垂直整合是综合服务能力的体现。不具备垂直整合能力的公司，将在竞争中处于不利的地位。

本节重点介绍信息产业内部垂直整合的趋势，开放源代码运动客观上加剧了信息产业上游的同质化，增加了产业链下游企业应用软件厂商的博弈能力。

开源软件加剧信息产业基础软件同质化趋势

开源软件的兴盛和发展，是推动信息产业不断前进的不竭动力之一。开源软件是送给中国信息业的一份大礼，是国内软件公司对抗微软等巨擘的强有力的武器，是国家信息安全的保障，也是促使信息产业同质化的最重要的因素，成为应用软件厂商向上游扩张的王牌。但是多年以来，中国公司对待开源软件的态度非常奇怪，首鼠两端，国家似乎也没有形成一致的意见。

软件公司既羡慕又妒忌微软的软件霸权，于是自主版权的呼声高涨，在民族主义和国家安全的名义下，搞了许多空费国家钱财而毫无收获的事情。对自主版权的强调和渴望，驱使某些短视的公司，把开源软件改头换面，就号称自主版权，形成事实上的“窃取”行为。那些对开源社区贡献最大的公司，反而是最优秀的公司。譬如，华为在 Hadoop（开源的大数据基础软件）社区重要贡献公司名单排名第七，是贡献最大的中国公司。而那些从不反馈开源社区的公司，并没搞出什么惊人的成就。

曾有一家创业型的公司来咨询如何通过大数据概念吸引投资。细问之下，这家公司就是照搬开源软件，给一些要求不高的客户实施数据仓库类项目。如果提供开源软件的服务能帮助客户解决具体的业务问题，笔者会举双手欢迎。但是一方面巧立各种名目，一方面拿开源软件来充数，此行为与诈骗无异，既不利于开源软件本身的发展，也不利于公司积累核心竞争力，最终竹篮打水一场空。

大胆地拥抱开源软件，充分地反哺开源软件，是中国软件企业在基础软件领域反超的唯一可行道路。这是一个开放的世界，在源代码领域也是一样的。学术圈和产业界都有那些两边“偷盗”的人，偷开源软件的果实，骗取国家的补贴，同时还打着民族和自主版权的幌子，指望他们是无法振兴信息产业的。因为他们一定是目光短浅的，一定是固步自封的。之所以从来没有对开源有贡献，是因为他们从没有超越开源软件的功能。

华为对于开源的态度和做法值得大力提倡。第一，华为明确宣称自己就是在使用开源软件，同时加入多个开源组织；第二，华为基于开源软件，开发商业应用，超越了开源软件所提供的功能；第三，华为同时将部分技术反馈给开源社区，反过来促进开源软件的发展。

开源软件的兴起最早可以追溯到 1955 年，一些年轻人为了深入研究 IBM 的操作系统，及时交换编程资料，成立“IBM USER GROUP SHARE”小组。Linux 开源操作系统的诞生，是开源软件发展史上的一个重大里程碑事件。它是 Unix 操作系统的开源实现和超越，最初是芬兰赫尔辛基市的一个天才大学生发布的，他名叫林纳斯 • 本纳第克特 · 托瓦兹(Linus Benedict Torvalds)。

Linux 是一款免费的操作系统，用户可以通过网络或其他途径免费获得，并可以任意修改其源代码。这是其他的操作系统所做不到的。正是由于这一点，来自全世界的无数程序员参与了 Linux 的修改、编写工作，程序员可以根据自己的兴趣和灵感对其进行改变，这让 Linux 吸收了无数程序员的精华，不断壮大。现在大名鼎鼎的安卓操作系统（谷歌推出的开

林纳斯 • 本纳第克特 • 托瓦兹(Linus Benedict Torvalds, 1969—)，著名的电脑程序员、黑客，Linux 内核的发明人及该计划的合作者。托瓦兹利用个人时间及器材创造出了这套属于当今全球最流行的操作系统（作业系统）内核之一。



“有些人生来就具有统率百万人的领袖风范；另一些人则是为写出颠覆世界的软件而生。唯一一个能同时做到这两者的人，就是托瓦兹。”美国《时代》周刊对“Linux 之父”林纳斯 • 托瓦兹（Linus Torvalds）给出了极高的评价。甚至，在《时代》周刊根据读者投票评选出的二十世纪 100 位最重要人物中，林纳斯居然排到了第 15 位，而从 20 世纪的最后几年就开始霸占全球首富称号的盖茨不过才是第 17 位。

——百度百科

源智能手机操作系统）也是从 Linux 修改而来。

开源软件客观上加剧了基础软件市场同质化的趋势。几乎每一款成熟的商业应用软件，都有对应的数款开源软件，见表 6-1。

表 6-1 部分进入商业主流应用的开源软件

	类型	优秀开源软件	商业软件
主流应用程序	搜索引擎	Apache Solr	Bai du 百度
	内容管理系统	Drupal	SMART ECM
	ERP 系统	openbravo	用友 yonyou
	商业智能套件	pentaho	ORACLE 甲骨文
	CRM 系统	SUGARCRM.	ORACLE 甲骨文
桌面系统及移动软件	压缩/解压缩	7ZIP	winrar
	移动 OS		
	浏览器	chrome	
	办公软件	LibreOffice The Document Foundation	Office
	虚拟机	VirtualBox	vmware
	PDF 工具	PDFCreator	Adobe Acrobat
	通讯工具	pidgin	msn
	播放器		real
操作系统与处理工具	操作系统	Linux	Windows UNIX
	数据库	MySQL	ORACLE 甲骨文 IBM DB2
	技术工具	hadoop	

几乎所有的大型平台级的互联网公司，其网站的架构都是以开源软件为主。谷

歌公司在发展的初期，因为缺少资金，无法购买昂贵商业服务器，不得不买一些淘汰的服务器，然后使用开源的 Linux 操作系统。二手服务器出现硬件故障的机率比较高，尤其是硬盘等存储设备，一旦损坏，数据就会丢失。不得已，谷歌公司自己开发 GFS 文件系统，解决硬件故障导致的数据丢失问题，成功地发展出分布式存储、访问技术。雅虎公司的一个开源小组，在谷歌成就的基础上开发出 Hadoop，这正是目前大数据技术领域最热门的方向之一。

开源软件，是送给信息产业的一份厚礼。那些善于使用开源软件的公司，将获得向产业上游扩张的技术能力，但同时需要反哺开源社区，这也是促使自己不断进步的手段。

大数据领域开源技术发展如火如荼，目前所有商用的号称提供大数据处理能力的一体机也罢、解决方案也罢，都集成了开源软件。但是这个领域还没有诞生一家有实力的提供开源技术服务的公司，就像 Red Hat 公司支持 Linux 发展一样。

Red Hat 是全球最大的开源技术厂家，其产品 Red Hat Linux 也是全世界应用最广泛的 Linux。Red Hat 公司总部位于美国北卡罗来纳州，在全球拥有 22 个分部。Red Hat Linux 操作系统盈利的主要来源是收取技术支持的费用。公司也销售收费的 Linux 系统，但是相比微软的 Windows，Red Hat 操作系统是开放源代码的。根据雅虎财经数据显示，Red Hat 目前市值接近 100 亿美元。

同样的，围绕 Hadoop 系列开源软件，也可能会产生一家主导型的开源技术公司，推动发展 hadoop、mapreduce、storm 等最新的数据处理技术，最终形成有竞争力的解决方案。

企业信息化市场垂直整合趋势

信息产业近几年垂直整合的风潮愈演愈烈。计算机缔造者之一 IBM 公司，一直以来都能够给客户提供从存储、主机、操作系统、数据库、中间件、应用软件的完

整解决方案，是不折不扣的蓝色巨人。甲骨文（Oracle）公司在强人拉里 · 埃里森的带领下，首先在数据库软件市场站稳脚跟，随即向应用软件市场进军，自主研发加一系列令人眼花缭乱的收购，已经成为全球第二大应用软件提供商，在全球企业管理软件市场仅次于 SAP 公司。但甲骨文并没有停下收购的脚步，开始利用庞大的客户群优势，向产业链上游进军，大手笔地收购了 Sun——一家 UNIX 主机厂商。众所周知的微软公司，在操作系统领域奠定霸主地位后，立即向产业链下游扩展，推出数据库产品，收购小型应用软件厂商，提供企业管理服务。

海外这些行业巨擘中，甲骨文公司的成长历程最具代表性。国内垂直整合信息产业链的是华为，华为的一小步已经代表中国整个信息产业的一大步。德国软件巨头 SAP，也是中国本土公司用友软件最大的竞争对手，已经收购了一家数据库公司，迈出向上游垂直整合的坚实步伐。用友软件在这一波大潮中如何迎头赶上，是产业界和资本市场非常关注的一件事情。用友或许会效法甲骨文和 SAP，向产业的上游进军，如图 6-3 所示。

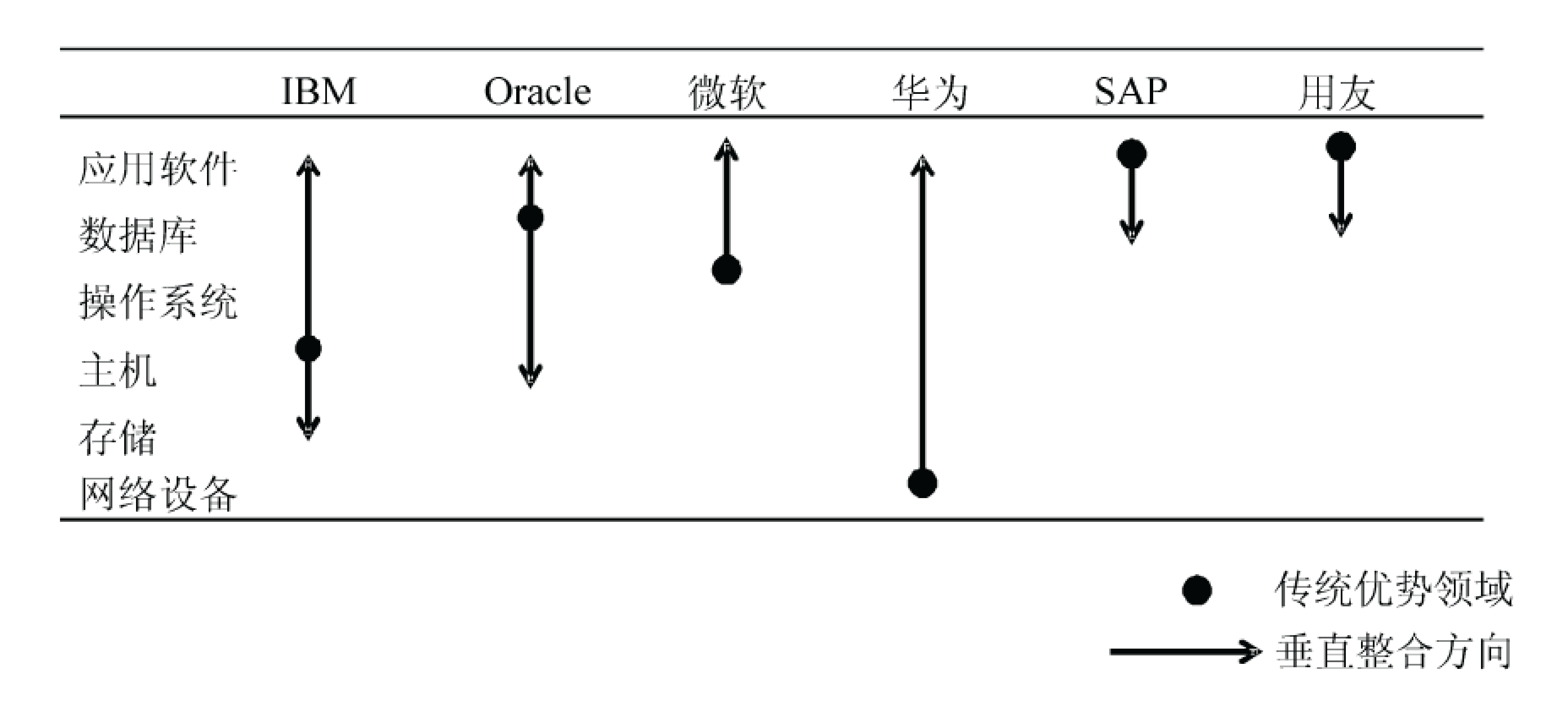


图 6-3　企业 IT 领域，行业垂直整合的趋势对比

甲骨文公司在拉里 • 埃里森的领导下极具进攻性。甲骨文开放平台数据库领域独占鳌头后，首先向下游扩张，横扫软件领域。2004 年收购了企业人力资源管理软件同时也是其竞争对手的厂商 Peoplesoft；2005 年收购全球最大的 CRM 软件厂

商 Siebel，使之成世界第一的 CRM 应用软件提供商；2007 年收购商业智能分析（BI）厂商 Hyperion（海波龙），加强对终端客户的掌控，直接为其客户提供应用软件、咨询服务，成为与德国软件巨头 SAP 分庭抗礼的企业管理软件供应商；2008 年收购项目组合以及管理软件的供应商 Primavera 软件公司，并在 09 年给项目管理软件产品升级，同时命名为 Oracle Primavera。

接下来甲骨文公司又向产业链上游扩张，打造全方位服务能力。在基础软件领域，2008 年收购了中间件巨头 BEA，使中间件市场大洗牌，挤压了大量中间件厂商的生存空间。其后，甲骨文插上了硬件的翅膀，2009 年收购了与自身具有强大互补性的操作系统、硬件平台厂商 Sun。Sun 拥有 SPARC 处理器和 Solaris 操作系统；同年，完成对虚拟化产品商 Virtual Iron 的收购。这些使甲骨文补齐了短板，形成了与 IBM 一样的从硬件到应用层全部涉及的 IT 巨头，引发了当时“红色巨人 PK 蓝色巨人”的激烈讨论。 甲骨文通过收购完成行业的垂直整合，股价一路走高，如图 6-4 所示。

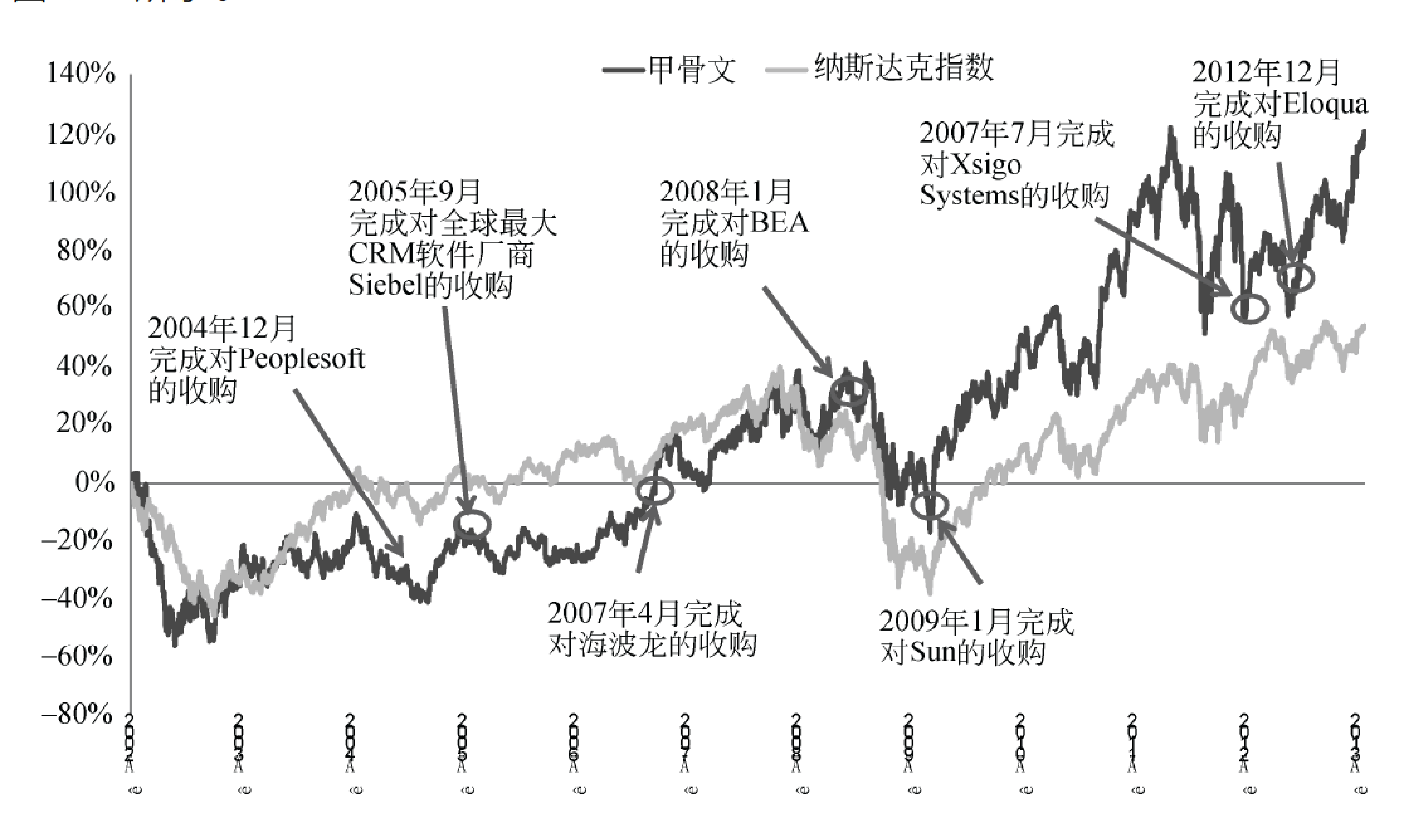


图 6-4　甲骨文公司沿着产业链垂直整合，推动公司股价和业绩持续增长

在收购了 Sun 后，甲骨文发布了一款集 Sun 硬件和甲骨文软件于一体的新数据库云服务器 ExaData，提供数据仓库和商务智能类系统、OLTP 类系统、混合负载类系统、数据库云平台服务。甲骨文不仅仅是将服务器、存储、IO 和虚拟化软件集成在一起，更是在于其对数据库、中间件和应用软件的深刻理解。

甲骨文公司 2012 年第三季度财报显示，甲骨文软硬一体化集成设计系统的硬件收入在该财季增长了 139%，是甲骨文历史上成长最快的产品。

华为公司垂直整合的思路最为清晰、坚决。IBM 曾经是华为倾心学习的老师，现在则是其最大的潜在威胁。华为的产业链甚至比 IBM 还要多出一层“网络设备”。应用软件尽管没有 IBM 完备，但华为通过“被集成”的策略和国内多家应用软件供应商合作，将产业上下游的产品集成在一起，为客户提供完整的解决方案。

当公司大到像华为一样的时候，就必须在产业战略层面思考公司是否安全的问题，思考是否可能被产业链上游或者下游的公司扼住咽喉要道的大事情。道理很简单，华为设备已经威胁到竞争对手，人家当然要从操作系统、CPU 等领域来遏制华为的增长势头。从这个意义来说，垂直一体化，是所有大型公司不得不走的一条路。

第三节　产品层面软硬一体化重获青睐

提要：

1. 软硬一体化是行业垂直整合趋势的具体表征和产品层面的载体。无论是消费电子市场，还是企业应用市场，都将呈现软硬一体化的特点。
2. 软硬一体化符合客户追求使用体验，要求运营维护简单的核心诉求。
3. 消费电子领域是小米公司，在企业服务市场华为公司是中国软硬一体化的代表。在海外大型的 IT 服务商，无一例外都推出大型的软硬一体化产品。

具体到产品层面，信息行业内的垂直整合则表现为软件、硬件一体化的趋势。苹果的 iPhone 是整合了包括芯片、主板、外壳、操作系统、主要应用软件在内的完整产品。甲骨文（全球最大的数据库提供商、企业管理软件供应商之一）公司桀骜不驯的创始人拉里•埃里森是乔布斯的好朋友，丝毫不掩饰对苹果软硬一体化设计思想的推崇和喜爱，在企业信息服务市场，邯郸学步般推出 ExaData 一体机，广受欢迎。软硬一体化，在苹果和甲骨文两大公司的引领下蔚为大观。

个人计算机时代，没有一家 PC 生产商的市值超越为个人计算机提供操作系统的微软。历史往往惊人地相似，但绝不会简单重复。我们以苹果、三星、诺基亚、索尼四家智能手机制造商为例，苹果的一体化程度最高，三星次之。诺基亚和索尼仅仅能够主导最终的产品设计，软件、核心硬件（处理器、内存、显示屏）都需要集成第三方的产品。诺基亚绑定微软公司，索尼拥抱谷歌公司。目前的竞争态势，明显是苹果和三星占据了上风。连微软——一直固守在操作系统领域的软件巨擘，也开始生产软硬一体化的智能手机和平板电脑，向苹果的 iPhone、iPad 正面宣战。可以理解成这是微软向软硬一体化趋势妥协的举动，如图 6-5 所示。

	(Apple)	SAMSUNG	NOKIA	SONY
核心业务	终端销售	终端销售	终端销售	终端销售
软件	自主	采购	采购	采购
硬件（处理器，内存，显示屏）	半自主	半自主	采购	采购
终端设计	自主	自主	自主	自主
	一体化凸显竞争优势		关键部件受制于人	

图 6-5　高度的软硬一体化竞争优势明显

业界软硬一体化的潮流，有深刻的产业背景。事实上，对于最终用户而言，只

要关心是否满足自己的需要，是否满足业务要求即可。具体是硬件实现还是软件实现，是自主也好，集成也罢，最终用户并没有那么关心。但有两点特别重要，一是大道至简，二是体验为王。

大道至简

唐朝大诗人白居易每写一首诗就读给老妇人听，老人如果听不懂就继续修改。“诗以载道”，白居易希望诗歌肩负教化社会的重任，所以他追求通俗浅近，能让更多的人理解诗中的寓意。正所谓“简则易知，易则易从”，一个产品若想广受欢迎，也必须符合易知、易从的原则。苹果手机正面只有一个按键，谷歌、百度的页面只有一个大大的搜索框，他们都是践行大道至简的典范。

计算机操作系统的发展历史上有两个重要的里程碑，先后成就了两家伟大的公司。第一个是由字符界面的磁盘操作系统（DOS）向图形界面的视窗操作系统（Windows）过渡；第二个就是丢弃了鼠标键盘，以手指触摸操作为主的 iOS。前者是大名鼎鼎的微软，后者是如日中天的苹果。这两次革新，都是大大简化了计算机的操作，让更多的普通用户，可以感受科技魅力，都把用户群扩大了数倍。

DOS 时代，操作计算机、编辑文档需要熟知许多命令。比如复制文件，就需要记住“copy”这个单词。如果想要搞点花样，还需要分辨“xcopy”和“copy”的区别，弄清楚一大堆参数的意义。在 Windows 中，这一切都不复存在，只需要用鼠标把文件“拖”过去就行。这个操作方式上的改变，把可以使用计算机的人数放大了成千上万倍，一下子把个人计算机推向办公主流地位。

iOS 的出现又一次大大简化了操作，进一步降低了计算机的使用门槛。触摸操作非常符合人们的天性，非常自然。即便是尚未识字的少儿和目不识丁的老人，都可以使用平板电脑玩游戏，进一步扩大了平板电脑的潜在客户。根据苹果公司披露的运营数据，iPad 平板电脑自 2010 年推出以来，在短短的两年内销量已经超过 5500 万台，截止到 2012 年 6 月，运行 iOS 系统的设备销售了 4.1 亿台。

在企业信息化市场，最近几年，企业用户并没有感受到信息产业的进步带给他们革命性的变化。首席信息官们不得不面对众多硬件厂商、软件厂商、系统集成商，终日陷于各类层出不穷的问题之中。悲剧的是，出了问题往往不容易定位引发问题的原因所在。硬件厂商说自己的设备一切正常，软件厂商说自己的软件日志中没有记录任何异常数据，互相推托指责。

任何一个应用系统都需要一批维护人员，有人专门负责网络，有人专门负责主机、存储等硬件设备，还有人专门研究数据库，当然更少不了应用软件的维护人员。机构日益臃肿，但问题越来越多。

现在 A 股市场已经有专门帮客户做信息系统维护的上市公司。把企业客户从繁琐的系统运行维护中解脱出来，专心于其业务发展，这应是 IT 产业发展的方向。企业应用的复杂性可能让这种主意听起来像痴人说梦。但是，IT 的复杂性的确不是客户需要关注的问题，是原厂商应该努力解决的。

造成企业应用复杂现状的根源，其实是由信息产业现有的格局和分工决定的。IT 投资黑洞、IT 产业分工，已经成为制约企业进一步普及 IT 应用的限制性因素。谁能率先打破既有产业格局，真正简化用户使用、维护的难题，谁就能像苹果公司一样，在企业服务市场笑傲江湖。

大道至简，并不是简单的消减功能，而是以最终用户和消费者为中心，高度抽象提炼其业务，把与业务无关的细节、复杂性全部隐藏在简洁的用户交互“界面”背后。打破硬件和软件之间的界限，以简化用户操作、降低用户维护成本，使用户专注于业务为最高目标。这需要深刻的行业洞察和强大的软件、硬件集成能力。

在企业 IT 基础设施领域，需要一到两家具备完整产品线的公司。在应用软件领域，则需要形成两到三家的供应商。在实施领域，则需要大型的集成公司为客户提供完整的端到端的解决方案。当这种一体化的设计能够真实地改善客户的业务、降低维护的复杂度、让客户把精力聚焦在如何更好地开展业务上，而不是无休止地处理 IT 引发的种种故障时，将是企业 IT 的一次飞跃。

体验为王

自从亚当•斯密开创性地描写了一枚大头针的制造过程后，工业社会就浩浩荡荡开始了产业分工的浪潮。一枚小小的大头针被分解成 18 个工种，有的工人专门负责拉丝，有的人专门负责抛光，有的人专门负责安装圆头……在艺术的殿堂，从未听说画家作画时会让不同的人给他画山川，另外的人来画河流，或者请人打打底色等事情。顶级奢侈品总是强调百年传承手工工艺、独一无二的材质、与众不同的感受。

工业社会确凿无疑地让各种各样的商品充斥在大街小巷，但是逡巡回顾之中，也难以发觉让人眼睛一亮的东西。所以艺术带来的精神享受变得弥足珍贵。传统的手工艺人开始得到联合国的救助，命名为“文化遗产”，希望能得到继承和发扬。

苹果是一家特立独行的公司，如其创始人和精神领袖斯蒂夫•乔布斯所言，“我一直站在科学和人文的交叉点”。苹果公司的确把工业设计和艺术人文结合在一起，给消费者一种流畅、享受的精神愉悦，形成极致的用户体验。在乔布斯眼中，似乎没有硬件、软件的区别，也没有制造产业、互联网产业、音乐产业等的区隔，只有消费者愉悦的快感。在苹果最低谷的时候，他们也没有放弃这种追求，反而不断强化其“端到端”软硬一体化的设计，推出颠覆性的产品，最终登上了全球市值第一的宝座。

对于企业用户而言，在业务高峰的时候，管用，就是最好的体验。许多行业都经历了业务量的爆炸式增长给系统稳定运行带来极大的压力。铁道部 12306 订票系统，饱受诟病的主要原因就是性能太差。对于企业用户而言，性能就是体验。

现在甲骨文公司、SAP 公司，在宣传自己的产品时，无不把快速处理当做最突出的特征。比如甲骨文宣传 ExaData 一体机，处理数据仓库类应用，速度比以前快 10 ~ 100 倍，联系事务处理的速度比以前快 20 倍。

软硬一体化的小米手机

雷军被其崇拜者称为“雷布斯”，意思是最像苹果的创始人乔布斯，推崇软硬一

体化的设计。小米公司 2010 年成立，2011 年推出小米 1 代手机，2012 年推出小米 2 代手机，通过网络预定来销售。小米 M2 手机正式开放网络发售时，首轮 5 万部在 2 分 51 秒内被抢购一空。2012 年 6 月达成的一笔融资中，小米公司估值是 40 亿美元。

小米公司毫不讳言是苹果公司的模仿者。不同于“山寨”厂商，小米有着自己的一套“模仿”理论——铁人三项：必须在硬件、软件、移动互联服务方面，都要处于领先位置，就像运动员要一气呵成“游泳、公路自行车、长跑”这三项极耗体力的运动一样。

在其短短的发展史上，已经发布了 1 代、2 代两款手机。软件产品包括 MIUI（米柚）操作系统、米聊、小米读书、小米分享、小米便签等。米柚操作系统根据 Android 系统深度定制，米聊、小米读书将承载不同的互联网服务。

在小米的成长之路上，故障、责难与质疑一直不断。但是同时掌控硬件、软件，为客户提供端到端的内容服务，无疑是一条正确的路。虽然荆刺丛生，但其未来和前景，也同样不可限量。

没有硬件这个躯壳，再好的软件也失去依托；缺少软件这个灵魂，再好硬件也没有生命。同时掌控硬件、软件，将其完美地融合，才能给用户提供完美的体验。个人计算机时代，不同厂商生产的台式机也好、笔记本也罢，无论怎么在硬件上推陈出新，但是用户打开电脑看到是都是 Windows 的操作系统，千篇一律。标准化的硬件制造商们，沦落到同质化的竞争，赚取微薄的利润。缺少掌控核心软件，只能沦落到为操作系统厂商打工的命运。

没有自己的硬件，哪家重量级合作伙伴愿意为小米内置关键的软件应用呢？譬如米聊、小米读书、网盘等。华为生产的手机中，一定是内置华为网盘；联想手机同样把联想网盘放在关键地位。因此，没有硬件，就会被竞争对手扼住喉咙，无法培育自己的移动互联网服务。看看苹果公司现在的动作，就更能体会软硬一体化的优势。很多原先排在 iPhone 首屏的图标，都一个一个地替换成苹果自家的软件。

小米 2 代手机的热销，已经显示其阶段性的成功。这种软硬一体化的思想，不仅仅适用于手机的生产。放眼各类智能型消费电子产品，无不遵循这一模式。小米完全有可能把小米手机的经验复制到其他的电子产品上：小米电视、小米平板等。缺少这种能力的公司，在扩张产品线时，必将受制于人。

企业市场的软硬一体化

另外，中国 IT 产业也一直存在怪圈：通用的硬件产品，在日益激烈的竞争中，价格趋于透明。大家并不认为国产的硬件设备有什么过人之处，尽管产业界做出了惊人的努力，但是客户的理由似乎更加充分，CPU、操作系统和硬盘都是国外的产品，硬件的技术含量如何体现呢？单纯的硬件厂商不得不滑向价格竞争的泥淖。

而应用软件在客户投资中，一向只占很少的一部分。客户的理由也很简单，几张光盘能值多少钱呢？对于实施服务，客户就用简单的数数人头的方式，来决定投资的金额。所以在现有的信息产业格局中，中国的硬件厂商和软件厂商都处于不利地位。

当下的信息产业格局，既不利于客户专注业务发展，又制约中国软件、硬件厂商的长远投资。硬件供应商是瞎子，因为他不知道客户业务需求；软件供应商是瘸子，因为没有硬件支持。瞎子背瘸子，跌跌撞撞走到现在实属不易。中国最大的软件厂商用友的销售收入 40 亿出头（2011 年的数据），这个数字放到国际市场上实在是叨陪末座。

但是，当硬件和软件紧密集成在一起的时候，局面似乎朝向有利于信息产业的方向转化。软件带来的神奇功能，叠加到硬件硕大冷傲的躯体中时，客户会觉得“哇，这个家伙很强大！”就会多花些银子。笔者在很多领域都观察到这个现象，苹果公司一直坚持这种端到端的策略，苹果手机毛利率几乎超过 50%。笔者的一位朋友销售应急调度指挥的软件一直打不开局面；而当开始提供“应急值班工作台”的时候，局面豁然开朗。他只是把调度软件和电话机、电脑装到一个柜子里，再配上一张桌

子，一个令人眼前一亮的新型办公设备就诞生了。对用户而言，再也无需自己安装软件、解决电脑故障等等问题，只要恪守自责，根据软件的提示打打电话就好。这是一个双赢的例子，现在朋友的公司已经受到资本的关注。

在企业应用市场，同样可以观察到软硬一体化模式带来的变化，海外甲骨文公司最先引领这个风潮。拉里•埃里森是苹果传奇创始人史蒂夫•乔布斯的密友，《乔布斯传》中提到埃里森非常赞赏苹果的软硬一体化的思想，并准备大刀阔斧地用到企业市场。的确，软硬一体化是简化企业 IT 应用、运维的最新尝试。

甲骨文公司的 ExaData 一体机，融合甲骨文公司的商用软件产品，包括数据库软件、数据仓库软件、中间件等，以及 Sun 公司的主机平台（Sun 已经被甲骨文公司收购，但是依然销售 Sun 品牌的主机）。这是一个典型的软硬一体化的产品，存储服务器采用 Oracle Enterprise Linux 操作系统，包括开源 Apache Hadoop、Oracle NoSQL 数据库、Oracle 数据集成 Hadoop 应用适配器、Oracle Hadoop 装载器、open source Distribution of R、Oracle Linux 和 Oracle Java HotSpot 虚拟机。

IBM 的 Netezza 一体机将数据存储、数据库、数据处理、以及数据挖掘集成在一体机中。其中硬件部分分为磁盘仓、SMP 主机、Snippet Blade（S-Blade）和网络结构。SMP 主机是两台高性能的 Linux 服务器，两台服务器中一台是活动的，另外一台是备机。S-Blades 是智能的处理节点，每个 S-Blades 是一台独立的服务器，它包含了个一台 IBM 刀片服务器和一块 Netezza 特有的数据库加速卡。Netezza 的架构结合了 SMP(对称多处理)和 MPP（大规模并行处理）的优点，建立了一个能以极快的速度分析 PB 量级数据的设备。Netezza 系统将复杂的非 SQL 算法嵌入到 MPP 流的处理组件中，对庞大的数据量能够以“流水线”方式对复杂数据进行分析处理，消除将数据转移到单独硬件的延迟和开销，同时其性能也提高了几个数量级。

EMC 作为硬件厂商，一直以来都是存储方面的翘楚。但在硬件的附加价值低、

利润日渐微薄和 Oracle 等竞争对手“软硬兼施”趋势的夹击下，EMC 开始实行软硬一体化，加强大数据时代的实力。在大数据方面，EMC 布局已久。2008 年，EMC 收购了网络管理软件开发商 Smarts，以增强网络管理能力；2011 年收购了具有数据分析与挖掘能力的 Greenplum，进入了数据仓库/商业智能市场。Greenplum 能够交付超出传统数据库软件 10～100 倍的性能，是 Oracle、Teradata 和 Netezza 等老牌厂商的挑战者。2011 年 10 月，EMC 收购数据库优化公司 Zettapoint，2012 年收购具有灵活研发计算能力的公司 Pivotal Labs、IT 绩效管理软件供应商 Watch4Net。并购后的 EMC 不再是一个硬件厂商，其基于自身在存储方面的实力把硬件和软件整合在一起，通过数据存储，帮助企业有效地管理内部的数据资产，转型为能创造更高的商业价值的综合解决方案提供商。

EMC 在大数据方面主要提供存储和统一分析平台，在并购 Greenplum 后，EMC 推出了 Greenplum 统一分析平台。EMC Greenplum 是数据库云平台，EMC 又将 Greenplum Database、Greenplum HD 和 Greenplum Chrous 整合推出大数据 Greenplum 统一分析平台(UAP)，使整个组织能够协作改变数据使用方式。

回顾本章的内容，当大家以数据的视角审视产业变迁的规律时，也许会对公司的价值和未来走向有一个全新的判断。在一次与业界技术高手的聚会上，有人突然发问，用友下一步会收购谁？沿产业垂直整合趋势，笔者几乎本能地回答，一定要收购一家搞数据库的公司。席间多人反对，理由多是数据库技术复杂，与用友整合困难等操作层面的问题。后一用友高层至，大家同样发问。他不假思索地说，一定要收购数据库公司。大家方释然！

导读：

1. 泛互联网化揭示未来将是以互联网为中心的数据社会，不同形态、不同类型各类终端，如果不集成网络功能，如果不能带来鲜活的数据，其商业价值将会大打折扣。泛互联网化是积累数据资产，发挥数据资产价值的最佳范式。
2. 苹果、谷歌、亚马逊、Facebook 这四家引领世界科技潮流的公司，均符合泛互联范式，尽管它们的盈利来源不尽相同。印象笔记的迅速崛起，与其成功运用该范式紧密相关。
3. 泛互联范式强调“终端、平台、应用”协同，公司需要根据自己的优势选择盈利的主要来源，盈利来源不同，其商业模式也不同。其最终决定胜负的将是“终端”或者“应用”带来的数据流量，是在“平台”中逐渐积累而形成的“数据资产”。

第七章

泛互联网化是发挥大数据价值的最佳范式

软件或者终端的价值，是由其承载的数据的流量与活性决定的。

——笔者

未来联网功能将内置在所有软件、硬件设备之中，是其功能不可或缺的一部分。网络浏览器依然是大家查看网页的首选工具，但是越来越多的人选择专用的软件工具，如微信、QQ 和微博。尤其是当大家通过智能手机、平板电脑等便携设备上网的时候，还可以通过扫描二维码，直接访问某些新潮的网站。有些书籍于多年前就提出未来家用电器应该全部具备联网的功能，像电冰箱、微波炉之类。当时这些看起来还比较遥远，但现如今联网功能的电视机都已经走入人们的生活了，只是操作稍显繁琐。汽车联网亦是大势所趋，举例来说，当大家在手机上享受即时更新的地图导航服务时，汽车 4S 店依然不紧不慢的每年给用户升级一次车载导航地图，还要收取不菲的费用。这种模式必将被互联网服务所取代，汽车即将成为大型的移动联网终端，人们即使在驾车的时候，也可以从网上即时获取信息，如拥堵路段、加油站方位等。下载更新地图只是车载软件后台的一个基本功能，无需司机们关心。

未来网络是“泛在”的网络，人们可以通过任何软件、任何设备在任何地点和任何时间获取网络服务。移动互联、桌面互联、汽车互联等都是泛互联网化的一种表现形式，这种命名方式是根据联网硬件设备种类来划分的，更强调它们之间的差异，而没有抽取它们之间共同遵循的产业规律。本章忽略联网终端的外在形态，抽象出联网功能本质的特征，来探讨一种具备产业生命力的范式。当超越移动互联、桌面互联等概念后，未来的商业图景反而变得更加清晰明了。对于泛互联范式的思考，基于两点假设：第一，人们越来越需要便捷的个性化服务，而非标准化的应用软件；第二，人们需要的是信息，而非承载信息的设备。

终端、平台、应用，加上大数据资产，构成“三位加一体”的泛互联范式（后文简称范式，参见图 1-14）。终端在本范式语境下，包括个人计算机、平板电脑、智能手机、智能电视、汽车等硬件终端，也包括音乐和视频的播放软件、编辑软件以及 QQ 聊天软件等软件终端。平台有两方面的意义，具备其中任何一方面，均可称为平台。第一是合作伙伴间利益共享的机制，强调其承载商业模式的特性；第二是不同应用共享数据的技术架构，强调其承载不同合作伙伴提供的应用程序的特性。

在信息世界，这两个特性往往互为表里，商业模式通过技术手段来实现，所以用一个术语来指代。应用则是指满足用户某些需求的软件程序。无论是终端、平台还是应用，在使用或者运营过程中，均产生各种各样的数据，包括日志、用户生成的文档资料、付费购买信息等等。这些都被精心地收集起来，形成“大数据资产”。在后续的章节中，大家会了解到数据资产可以演绎出不同的商业模式。

终端最典型的特征是门户化，无论是硬件还是软件，都可能成为用户完成某类工作、获取某类服务的必备之物和必经之地。门户有排他性和唯一性的特点，如某个软件一旦具备了门户的特征，那它就基本走在赢者通吃的路上，甚至给第二名都留不下多少机会；再如智能手机成为大家随时随地听音乐的首选后，MP3 类的消费电子产品，也就寿终正寝了。

软件产品具备门户特征的前提条件是具备在多种硬件设备上运行的能力。譬如印象笔记，迅速在 Windows、Mac、Android、iOS 等各种设备上开发终端软件，无论用户使用什么设备，都能有一致的软件使用体验。如微博终端软件、谷歌搜索服务等等，都具备这个特点。

门户化的价值在于吸引足够多的用户、足够快的使用频次，为碎片化长尾应用奠定基础。

平台化是指能够承载相关产品、服务，或者是第三方产品、服务的机制。承载自有产品、服务的核心在于底层数据架构及技术架构的一致性和拓展性。承载第三方产品、服务的核心在于利益共享的商业模式。一旦完成平台化，就具备了给用户提供全面服务的能力。平台也即成为多方获利的机制。平台拥有者，获得制定游戏规则的权力。维护平台的繁荣是伟大公司的必然选择。在美国，互联网领域最新的四大平台是谷歌、苹果、Facebook 和亚马逊。在中国，腾讯、百度、阿里巴巴正在上演三国演义。

应用最具未来趋势的特征是碎片化。把原来大型臃肿的软件，拆分成多个独立

的功能组件，用户可以按需下载使用。最典型的例子就是苹果的 App Store，每个“碎片”完成一个小功能，聚合起来，就可以满足人们方方面面的需要。到 2012 年 10 月，苹果应用商店中有 70 万种不同的应用，下载量已经超过 300 亿次。

碎片化的最大价值在于破解了厂商提供标准化产品和用户需要个性化服务之间的矛盾。碎片化衍生出微支付，用户可以只花几元钱就买到很实用、很好玩的东西。如果一些大型应用软件通过碎片化方式提供，还可以显著降低用户的总体拥有成本。

碎片化应用是平台拥有者的主要盈利来源。是靠终端产品获取收入，还是干脆从数据里面淘金？这是不同行业、不同发展阶段需要仔细斟酌的问题。但是最基本的原则是清晰的，就是不能伤害用户的体验。苹果的主要收入来源是终端类产品，包括 iPhone、iPad 等，应用商店中碎片化的应用，仅仅是终端收入的小零头。腾讯的主要收入来源则是附加在 QQ 平台上的应用，如虚拟的服饰、道具等，QQ 聊天软件本身是免费的。谷歌公司虽然现在也开始卖终端，如 Nexus 手机和平板电脑，但本质上，谷歌是靠深入挖掘数据来盈利的。即便谷歌卖手机，也要比苹果便宜得多。

缺少终端，就失去了战略的主动权，很可能沦为别人平台上的一个碎片化应用。缺少平台，则难以做大，无法形成有效的产业协同效应、聚集效应；缺少碎片化应用，就无法满足用户多层次的需求，难以解决标准化产品和用户个性化服务间的矛盾。

泛互联范式比较抽象，后续小节将通过几个例子来详细阐述。首先从苹果应用商店模式的前身——iPod 音乐播放器开始介绍，便于大家理解伴随着智能手机 iPhone 同时诞生的应用商店的意义。然后介绍印象笔记，一款让大家随时随地记录信息的软件。开发印象笔记的公司，目前在资本市场的估值为 10 亿美元。最后通过对比印象笔记、微软办公软件、谷歌在线文档三类不同的产品，阐述泛互联网化带给软件产品在商业模式、软件架构等方面的改变。

第一节　苹果——终端崛起

提要：

1. 苹果的泛互联网化之路，开始于便携的音乐播放器 iPod。iPod 开创的商业模式具备了泛互联范式的三大特征：门户化、平台化、碎片化。
2. 苹果的应用商店建立了全新的商业模式，使应用软件有了便捷的线上发布渠道，聚集了大量的第三方开发人员，让苹果的智能手机能够用于各类生活和工作场景。
3. iCloud 是苹果开启大数据战略的标志事件，通过 iCloud 可以收集用户应用程序数据。苹果显然并未就此止步，在收集用户消费行为数据方面，推出一系列的举措，包括 GameCenter、Passport 等产品。
4. 未来智能手机通过收集的各类数据，将是最了解用户的“人”。

iPod 的丰碑

iPod 是便携音乐播放器发展史上的一座丰碑，至今仍无人超越。“苹果公司先开发了 iPod 还是先开发了 iTunes 软件？”这个问题恐怕连最资深的苹果粉丝也难以回答。

在 2000 年左右的美国，人们热衷于从 P2P 软件中下载音乐并刻录到 CD 上，但下载软件、刻录软件以及刻录机的操作具有一定的门槛，只有发烧级的音乐爱好者才会钻研如何使用这些东西。乔布斯从中看到了巨大的商机，他收购了音乐管理程序 Rio 的创业团队，并用他一贯苛刻的要求使得该产品变得更简单易用，使用户

体验更优，这款产品就是后来的 iTunes[①]。

有了 iTunes 之后，乔布斯希望能有一个和 iTunes 配套的产品，让用户更轻松地收听音乐，这样 iPod 才被创造出来。事实上是先有 iTunes，后有 iPod，这和许多读者的认识恐怕有所不同。iTunes 创立之初面临着“巧妇难为无米之炊”的困境，而当时的唱片公司日子也不好过，整天在一系列的盗版案件中垂死挣扎。乔布斯凭借其在好莱坞的创业经验和天才的商业头脑，说服了五大唱片公司向其提供数字音乐的销售权。乔布斯计划把每首歌曲的价格定为让人心动的 99 美分，唱片公司将从中抽取 70 美分。于是 iTunes 商店诞生了，“音乐公司能赢利，艺术家能赢利，苹果公司也能赢利，而用户也会有所收获”的“四盈”商业模式最终被确立起来。iTunes 商店在推出后的 6 天内就卖出了 100 万首歌曲，在第一年卖出了 7000 万首歌曲；2006 年 2 月，iTunes 商店卖出了第 10 亿首歌曲；2010 年 2 月，iTunes 商店卖出了第 100 亿首歌曲。

在“iPod + iTunes 商店”模式中，人们发现硬件、软件、内容（音乐）首次完美的结合在一起，形成最佳的客户体验。苹果通过大量的 iPod，控制了音乐发行的渠道，从而引起整个音乐产业的变革，如图 7-1 所示。

在这个模式中，iPod 作为一款独立的音乐播放设备，非常受人欢迎。同类的 MP3 播放器，跟 iPod 的相比就像廉价的山寨货。iPod 已成为人们收听音乐的首选，没有人在使用 iPod 的时候，还会使用其他播放器。iPod 客观上具备了音乐门户的特征。

iTunes 商店则构建了和唱片公司的商业模式，分成比例接近 7:3，唱片公司占大头。在 iTunes 商店中，唱片公司不用担心盗版的困扰。苹果公司更进一步，直接

① iTunes 是一款媒体播放器的应用程序，2001 年 1 月 10 日由苹果公司在旧金山的 Macworld Expo 推出，用来播放及管理数字音乐与视频文件，至今依然是管理苹果电脑最受欢迎的 iPod 的文件的主要工具。此外，iTunes 能连接到 iTunes Store（在有网络连接且苹果公司在当地有开放该服务的情况下），以便下载购买的数字音乐、音乐影片、电视节目、iPod 游戏、各种 Podcast 以及标准唱片。

和有才华的音乐人签约，他们可以跳过唱片公司，直接在 iTunes 商店中，发行他们的最新作品。苹果公司取代了唱片公司部分职能，同时通过 iTunes 商店获利的第三方也大大增加，iTunes 已成为一个广受欢迎的音乐发行平台。

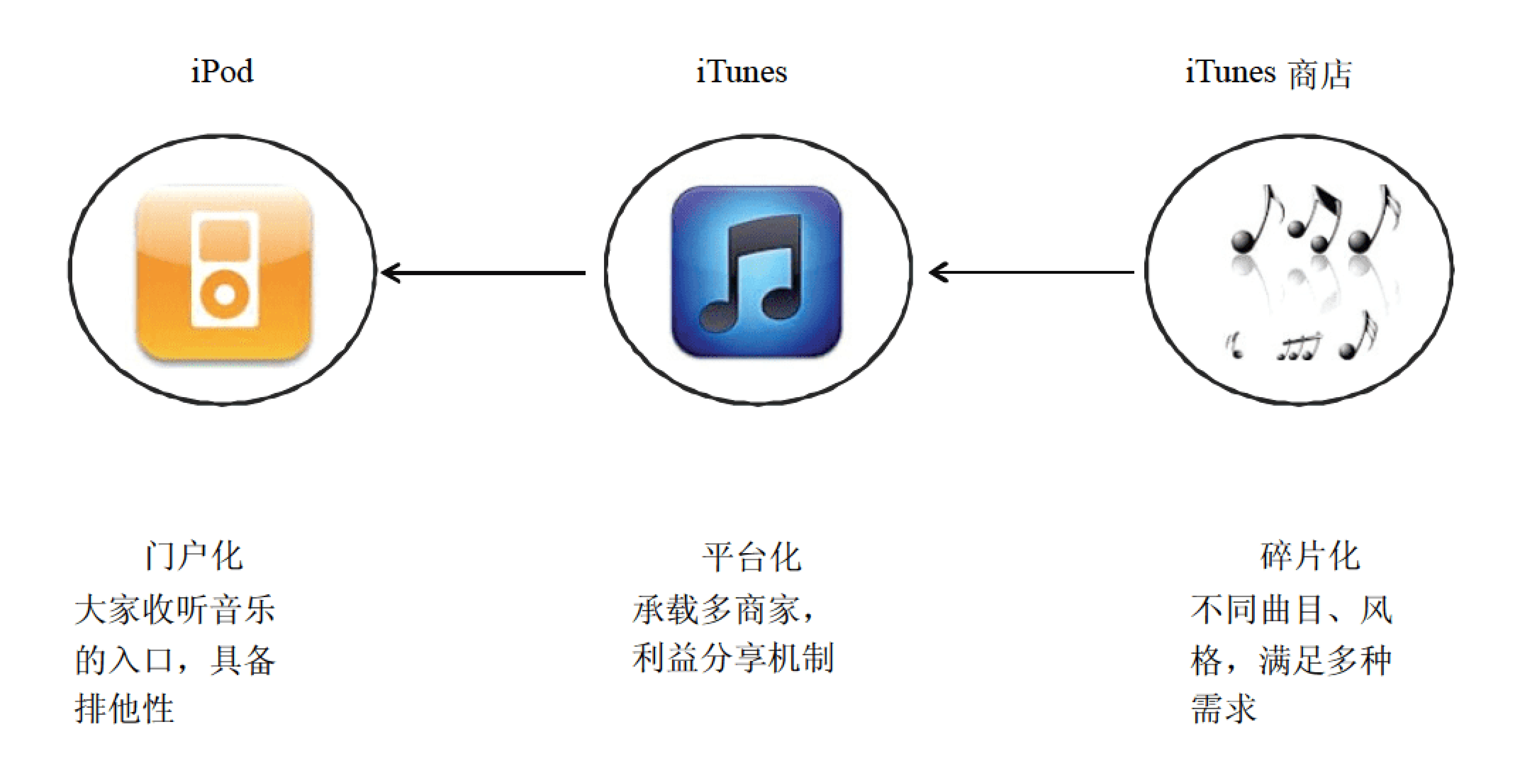


图 7-1　iTunes + iPod 开创了泛互联网化模式的雏形

消费者自然众口难调，苹果打破了按照唱片发行的惯例，用户可以购买单独的曲目，不再把好听的歌曲和差的歌曲混在一起强迫消费者购买。把唱片碎片化成单独歌曲，从而最大限度地满足了用户个性化的需求。

以消费者的立场，从数据的角度再来总结“iPod + iTunes”模式。音乐同时保存在 iPod 和 iTunes 中，这两者之间通过“同步”的机制来保持一致性。另外，同步的数据中还包括“播放列表”数据。播放列表就是消费者的“偏好”，极具个人色彩，你的播放列表和我的播放列表肯定是不一样的。在“iPod + iTunes”机制中，“播放列

App Store 即 application store，通常理解为应用商店。App Store 是一个由苹果公司为 iPhone 和 iPod Touch、iPad 以及 Mac 创建的服务，允许用户从 iTunes Store 或 Mac App store 浏览和下载一些为了 iPhone 或 Mac 开发的应用程序。用户可以购买或免费试用，让该应用程序直接下载到 iPhone 或 iPod Touch、iPad、Mac。其中包含游戏、日历、翻译程序、图库，以及许多实用的软件。App Store 在 iPhone 和 iPod Touch、iPad 以及 Mac 的应用程序商店都是相同的名称。

表”并不完全依赖 iPod，这就保证当人们换一个新 iPod 时，依然能够非常容易地找到自己喜爱的歌曲。

这种数据“同步”的机制，和纯粹的互联网应用是不同的。纯粹互联网应用在用户的“终端”是没有数据的。换句话说，泛互联网化的终端，在离线状态下，依然可以发挥核心的功能，如果在连线的状态下，则可以获得更多的数据。而纯粹的互联网应用在离线状态下，是不可用的。这也是泛互联网化应用在与互联网应用之间重要的差别。

应用商店打造全新的产业生态

iPod 非常成功，2005 年 iPod 设备的销售收入占据苹果公司收入的 45%。乔布斯不但没有志得意满，反倒深感担忧，他认为能抢走 iPod 风头的，一定是手机。当每部手机中都内置了音乐播放软件时，iPod 的路就走到了头。

幸运的是，苹果公司开发出了风靡世界的智能手机——iPhone。iPhone 的确如乔布斯所言，内置了 iPod 音乐播放器，不仅如此，还继承了 iPod 时代行之有效的“音乐商店”的做法，把音乐商店，扩展成“应用商店”。消费者可以通过应用商店下载各种各样有趣儿的应用软件，如给照片装饰一个相框，或者记录自己每天跑步的里程等等。

2008 年 3 月 6 日，苹果对外发布了针对 iPhone 的应用开发包，供免费下载，以便第三方应用开发人员开发针对 iPhone 及 Touch 的应用软件。3 月 12 日，仅用不到一周时间，苹果宣布已获得超过 100 000 次的下载；三个月后，这一数字上升至 250 000 次。众所周知，苹果公司一直以来在产品及技术上都具有一定的封闭性。在 IBM 推出兼容个人计算机之后，微软等一系列软件公司围绕 PC 开发了很多办公、娱乐软件，通过增强用户对软件的粘性争夺了很大一部分个人计算机用户。而苹果的 Mac 电脑由于其软件和硬件的兼容性问题一直不被苹果公司重视，因此只拥有 10%左右的“铁杆粉丝”。苹果这次推出 SDK 之举可以说是第一次向个人和企

业开发者抛出了橄榄枝。另外，用户购买应用所支付的费用由苹果与应用开发商按照 3：7 的比例分成，那些一战成名的暴富神话吸引了全球众多的企业开发者和个人开发者。在众多开发者众星捧月般的簇拥到 App Store 这个平台之后，一个商业生态系统悄悄地形成了。7 月 11 日苹果 App Store 正式上线，可供下载的应用已达 800 个，下载量达到 1000 万次。2009 年 1 月 16 日，数字刷新为逾 1.5 万个应用，超过 5 亿次下载。截至 2012 年 10 月，其应用数量已经突破 70 万，累计下载量也突破 300 亿。

“应用商店”催生了内容创造产业，其影响力波及整个信息行业，大家不约而同地在思考相同的问题，是成为苹果应用商店里一个碎片化应用，还是另起炉灶，创建自己的应用商店?

iPhone 作为最流行的手机之一，扮演“大门户”的角色，无论是打电话、玩游戏、刷微博还是阅读电子杂志，总是离不开 iPhone；应用商店扮演平台的角色，解决了与广大开发者之间的利益分配问题，并成为推广软件应用的主要渠道；应用商店里形形色色的各种碎片化应用，满足人们工作、娱乐、休闲、购物等多种需求，如图 7-2 所示。



图 7-2　App Store + iPhone 引发智能手机的变革，重新定义了产业生态，泛互联网化范式形成

再回到大数据的视角，来审视应用商店模式。用户只是在下载或者更新应用时，会使用应用商店，而用户使用应用程序而产生的“行为数据”和“内容数据”，并没有被收集和记录。换句话说，仅仅拥有消费者在应用商店中下载应用软件数据，还不足以构成“大数据”，这些数据的活性不足。

当 iPad 平板电脑推出后，数据问题就更加突出了。人们在 iPhone 中有大量的照片、通讯录、音乐、文档等等资料，但是如何方便地在 iPad 上看到呢？如果手机丢了，这些资料又如何找回呢？就这样，iCloud 应运而生了。

iCloud“个人数据中心”应运而生

2011 年 5 月 31 日，苹果公司官方发布 iCloud 产品，提供了邮件、日历和联络人的同步功能。除此之外，iCloud 还具有强大的存储功能，它可以存储人们购买的音乐、应用、电子书，并将其推送到所有匹配设备。可以说 iCloud 第一次使得包括 iPhone、iPod Touch、iPad，甚至是 Mac 电脑在内的所有苹果产品无缝连接，借助 iCloud 苹果也实现了从多个数据源收集数据并进行统一存储和索引的功能，为搭建大数据中心铺平了道路，如图 7-3 所示。

iCloud 具有以下几大功能：照片流、文档和应用云服务、日历、通讯录、邮件、iBooks 备份和恢复。我们发现，每个功能都是苹果收集用户数据的来源之一。

照片流。这一功能使得用户用一部 iOS 设备拍摄照片，影像就会出现在其他设备上，包括 Mac 或 PC。将照片从数码相机导入到电脑之中，iCloud 会即刻通过 WLAN 将它们发送到用户的 iPhone、iPad 和 iPod Touch 上。用户无需人为地去同步或是添加照片到电子邮件的附件中，也不必传输文件，照片就会出现在每一部苹果设备上。同时，用户可以选择指定的人群来共享照片。用户也可以让观众对照片发表评论，并可以回复他们的评论。照片流的功能使得用户影像数据得到统一保存，为影像数据的收集提供了方便。

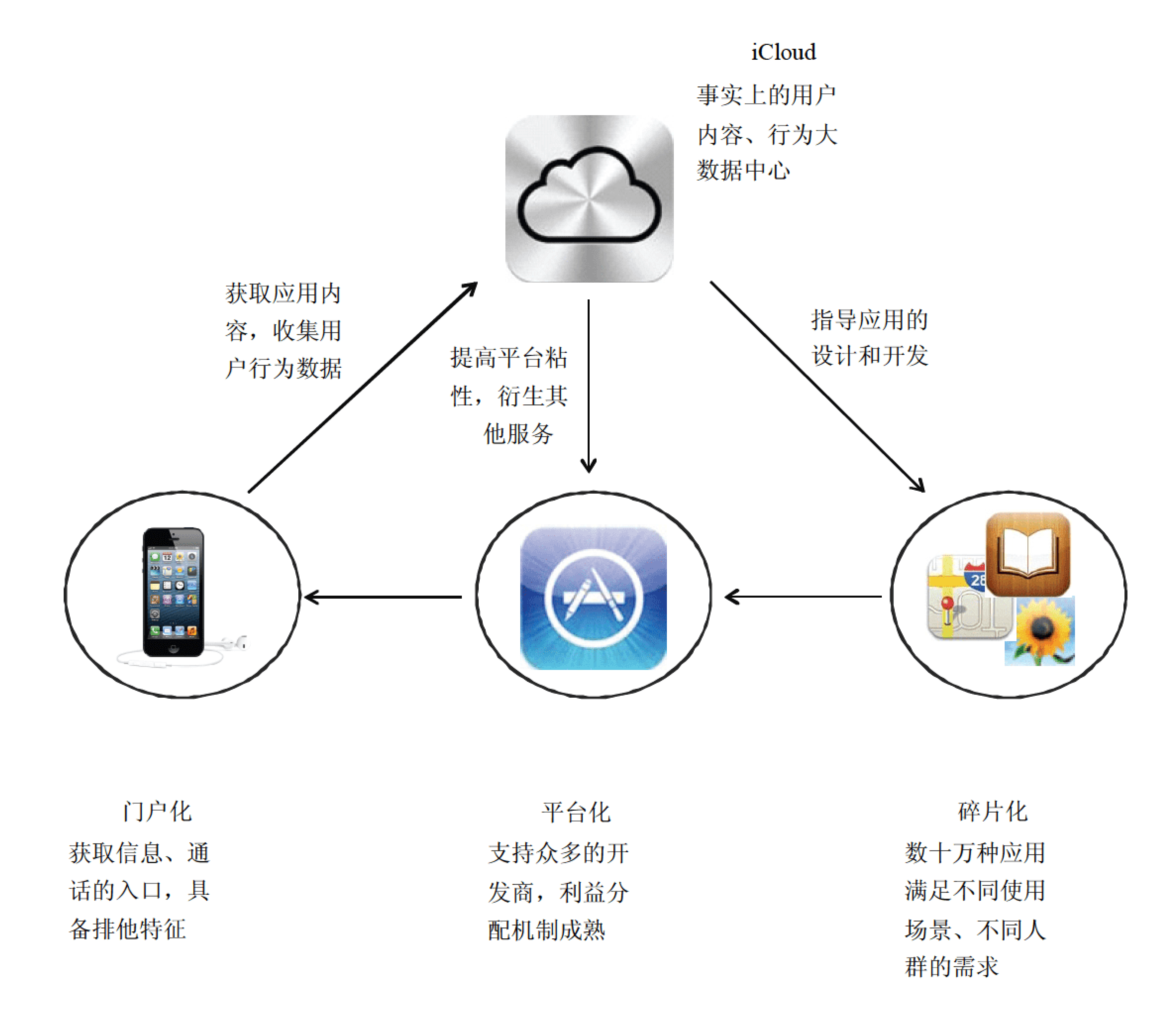


图 7-3　推出 iCloud，标志苹果完成泛互联范式的最后一块拼图

文档和应用云服务。用户可以在 Mac[①]、iPhone、iPad 和 iPodTouch 上创建文档和演示文稿。同样地，iCloud 可让该文件在 Mac 和所有 iOS[②]设备上保持更新。iCloud 已内置于 Keynote[③]、Pages[③]和 Numbers[③]等 App 中，此外还可与其他支持 iCloud 的 App 配合使用。同时，用户在某一设备上购买的应用也将自动同

① Mac：麦金塔电脑，俗称 Mac 机或苹果机，是苹果公司设计生产的个人台式电脑系列产品。

② iOS 是由苹果公司开发的操作系统，最初是设计给 iPhone 使用，后来陆续套用到 iPod Touch、iPad 以及 Apple TV 产品上。就像其基于的 Mac OS X 操作系统一样，它也是以 Darwin 为基础的。原本这个系统名为“iPhone OS”，直到 2010 年 6 月 7 日 WWDC 大会上宣布改名为“iOS”。

③ Keynote、Pages、Numbers 是适用于 Mac 的办公软件。Keynote 用于播放幻灯片，类似 PowerPoint；Pages 用于编辑文字，类似 Word；Numbers 用于处理电子表格，相当于 Excel。

步到其他设备中。这一功能具有革命性的意义，开发者通过苹果提供的 iCloud API[①]，可以将自己开发的应用产生的数据保存到云端。用户在使用这个支持 iCloud 的应用时，无需人为地上传或同步数据即可实现在多设备上同步地编辑文档。苹果公司也通过这种方式获得了更具价值的应用数据，进而为应用大数据打下了基础。

日历、通讯录和邮件。iCloud 可以存放用户的私人数据，包括日历、通讯录和电子邮件，并让它们在所有设备上随时更新。如果用户删除了一个电子邮件地址，添加了一个日历事件，或更新了通讯录，iCloud 会在各处同时做出这些更改。同样地，用户的备忘录、提醒事项和书签也会进行同步。日历、通讯录和邮件这三个数据源提供了用户最为私密的、也是价值最高的数据。苹果公司能够收集到用户的私人数据无疑会大大地提升其提供个性化服务的水平。

iBooks。由于移动阅读具有最为广泛的潜在客户群以及更为广阔的市场空间，因此其一直是各个终端厂商、服务提供商、应用开发商以及运营商争夺的地盘。苹果公司凭借其统一、流畅的用户体验赢得了众多用户。iCloud 出现无疑进一步巩固了 iBooks 的市场地位。一旦用户从 iBooks 获得了电子书，iCloud 会自动将其推送到用户的所有其他设备中。对于其他操作，iCloud 也会进行数据的同步。比如，用户在 iPad 上开始阅读，加亮某些文字，记录笔记，或添加书签，iCloud 就会自动更新用户的 iPhone 和 iPod Touch。

备份和恢复。用户的 iPhone、iPad 和 iPod Touch 上存放着各种各样的重要信息。在接通电源的情况下，iCloud 每天都会通过 WLAN 对它们进行自动备份，而用户却无需进行任何操作。当用户设置一部全新的 iOS 设备，或在原有的设备上恢复信息时，iCloud 云备份都可以担此重任。只要将设备接入 WLAN，再输入 Apple ID 和密码就行了。备份和恢复不仅是方便客户的功能，对苹果也极具意义，它最大化收集了用户的数据，可以衍生出其他服务，并指导应用设计和开发。

如果把时间切换到 60 年前，人们将发现 iCloud 的意义远远超过 iPhone 的成

①API（Application Programming Interface）又称为应用编程接口，就是软件系统不同组成部分衔接的约定。由于近年来软件的规模日益庞大，常常会需要把复杂的系统划分成小的组成部分，编程接口的设计十分重要。程序设计的实践中，编程接口的设计首先要使软件系统的职责得到合理划分。良好的接口设计可以降低系统各部分的相互依赖，提高组成单元的内聚性，降低组成单元间的耦合程度，进而提高系统的维护性和扩展性。

功。自计算机诞生以来，计算机一直扮演“数据中心”的角色。人们所有的文件、资料都保存在个人计算机中。iCloud 横空出世，将取代个人计算机的“数据中心”角色。iCloud 也不同于纯粹的互联网应用，其思想和 iPod 时代的音乐管理一脉相承，即泛互联网化。

“游戏中心”的战略意义

2010 年 9 月 9 日，苹果正式发布了 iOS 4.1，其中有一款具备战略意义的产品：“Game Center”游戏中心，如图 7-4 所示。



图 7-4　2010 年 9 月，苹果发布了 Game Center

游戏中心是专为游戏玩家设计的社交网络平台，Game Center 简化了兼容游戏中多人对战的配对流程。另外，它不但可以通过成就系统，同时也可以通过积分榜为玩家提供炫耀的资本。借助 Game Center，用户可以收发好友请求，可以邀请好友通过互联网参与多人游戏。除此之外，系统还可以自动为用户寻找游戏玩伴。用户可以在 Game Center 中看到游戏中的玩家排名和成绩，并且可以借助好友推荐来寻找新游戏，也可以直接进行聊天。

Game Center 一经推出就受到了用户的欢迎。考虑到 iPhone、iTouch 和 iPad 的操作性和屏幕尺寸，Game Center 的用户多数并非某大型游戏的玩家，他们可能更热衷于一些打发时间的休闲游戏，如疯狂的小鸟等。他们通过 Game Center 玩游

戏更多地是满足自己好奇炫耀的心理。苹果抓住了用户的使用心理，很好地为他们搭建了 Game Center 这个平台，并且借助 SNS 自我发展用户的特性吸引了大量的用户。在 2012 年 WWDC 大会上，苹果宣布其 Game Center 用户数量超过了 1.3 亿。Game Center 在短短两年时间内发展到如此规模的用户数，这在视频游戏领域是史无前例的，可以毫不夸张地说，Game Center 再一次改变了视频游戏的世界。

通过 Game Center，苹果获取了用户玩游戏的行为数据以及游戏社交数据，提高了平台粘性，从而构建起大型的娱乐消费数据中心，利用这些数据也衍生出了其他服务，如指导应用的设计和开发，如图 7-5 所示。

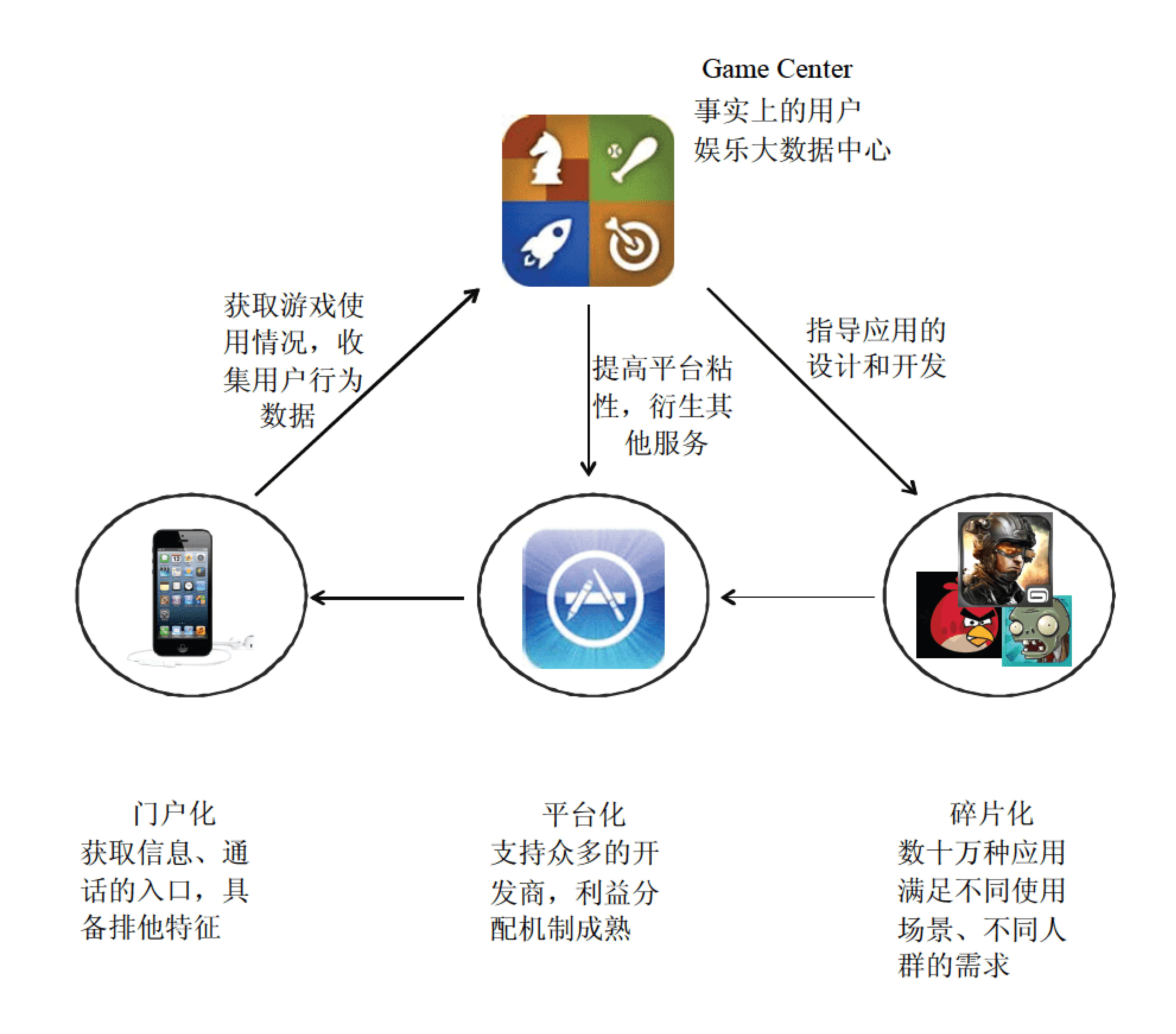


图 7-5　Game Center 成为苹果游戏领域的“数据中心”

总结上述三种模式，苹果公司完成了从应用商店的争夺到对用户行为数据的争夺，并通过三个阶段，苹果公司完成了构建消费者大数据中心的全部过程。

第二节　印象笔记（EverNote）的启示

提要：

1. 软件产业依然受到工业化时代标准化思维的影响，典型代表是微软的办公软件。功能全面但是臃肿的办公软件，并不满足泛互联范式。这类软件过去的辉煌，只能反衬其未来的没落。
2. 电子文件的结构和复杂性，增加了人们获取文件中蕴含的数据和信息的障碍。文件这个概念，或许应该被丢弃。数据和信息将以崭新的面貌被重组和分享，也许人们已经临近另一场信息革命的边缘。

EverNote 的中文名称是“印象笔记”，其主要功能是帮助大家快速地记录笔记，可以通过手写、键盘、录音、拍照等手段，类似于微软公司大名鼎鼎的 Office 系列办公软件中的 OneNote 笔记软件。对于用过这款软件的读者而言，对其易用性、多平台同步等特点肯定有所了解。对于大多数读者而言，一个笔记类的软件，又有什么出彩的地方呢？先看一组数据。

截止到 2012 年 6 月，印象笔记全球用户突破 3400 万，付费用户 140 万。而在 5 月，其注册用户数才 2500 万，付费用户数 100 万。印象笔记快速的发展速度令人震惊，照此发展，其营业收入很快就会突破 1 亿美元。资本市场慷慨地给这家公司估值 10 亿美元，这个估值水平超过绝大部分在 A 股创业板挂牌的公司的市值。

对比印象笔记、微软的 Office 办公软件和谷歌公司的 Google Docs 这三款产品，可以清晰地发现具备泛互联网化特征的产品蕴含的巨大商业价值。笔者依时间为序，首先分析微软的办公软件，接下来看看和微软一直对着干的谷歌，最后剖析印象笔记的商业模式和产业特征。

面向个人计算机的微软 Office 办公软件

微软的 Office 办公软件早在 1983 年就伴随着 MS-DOS 诞生了，随着微软系统版本的升级，Office 也同步升级到 95、Server、98、XP、2000、2003、2007、2012 等版本。很长一段时间里，用户安装完操作系统后第一件事情就是安装相应版本的 Office 办公软件。Office 办公软件已经成为用户办公、生活不可或缺的软件之一。因此很难讲是 Windows 系统成就了 Office 办公软件，还是 Office 办公软件带动了 Windows 系统的高市场占有率。

大家对微软的视窗操作系统印象最为深刻，但实际上，Office 系列办公软件，才是微软最大的摇钱树。在过去 3 年多的时间里，掌管 Office 办公软件的商业部门曾经在 10 个季度为微软贡献最高的利润。2012 年第一季度，微软总营收为 174 亿美元，商业部门营收的 58 亿美元，占到总营收的 33.4%。同期，Windows 和 Windows Live 部门营收 46 亿美元，仅占总营收的 26.6%。

历史上，有众多的竞争对手试图挑战微软在办公软件领域的统治地位，包括 IBM、Sun 等巨擘，但皆无功而返。微软的办公软件，凭借丰富的功能、正确的市场策略，占据了垄断地位。

可以把文档当成一种计算机交互的“语言”，把办公软件当成人们分享文档时的“翻译”，这样就非常容易理解这类软件具有天然的垄断优势。因为随着更多人使用这种“语言”，就会自然而然地离不开翻译，形成正向反馈。在如图 7-6 所示的循环中，使用的人越多，产生的文档就越多，文档越多，需要使用相同办公软件的人就越多。微软公司是靠卖软件的“拷贝”盈利的，每个人都必须购买一份“拷贝”。所以，可以观察到微软办公软件的功能越来越多，支持的文档类型也越来越多。在发展初期，增加功能和支持更多类型的文档，的确增强了不同办公软件之间的交叉销售能力。微软几十年来垄断了办公软件市场，现在微软办公软件功能之强大，远非竞争对手的产品可比。下面简单回顾一下微软办公软件的发展历程。

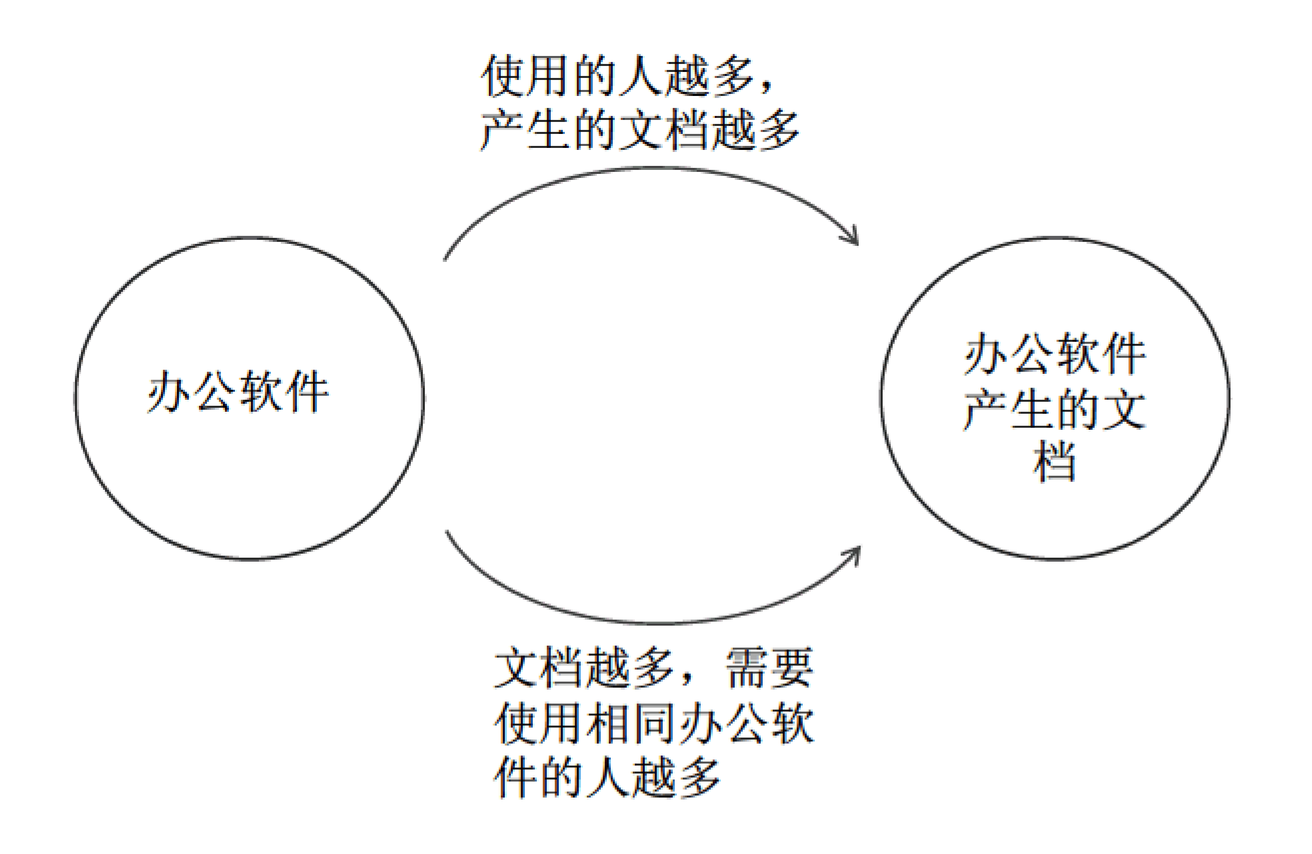


图 7-6　办公软件和文档相互促进

案例：微软办公软件发展简介

Microsoft Office 3.0 是第一版针对视窗系统所发布的办公软件。这个版本于 1992 年发布，包括 Word（字处理软件）、Excel（电子表格软件）、PowerPoint（演示文稿软件）三个主要软件，自此有了“办公套件”的概念，具有一定的历史意义。

Microsoft Office 4.0 于 1994 年推出，增加了 Mail（邮件软件）和 Access（数据库软件）。

Microsoft Office 4.3 是最后一版 16 位的版本，同时也是最后一版支持 Windows 3.x、Windows NT 3.1 和 Windows NT 3.5 的版本。

Microsoft Office 97 是一个重大的里程碑。这一版中包含了许多新功能和改进，其同时也引入命令栏（Command Bars）的功能以及拼写检查的功能。

Office 的以上各个版本均被称作“办公套件（Office suite）”，顾名思义，以上各个版本 Office 将办公常用功能软件打包，一并销售。基本上说，用户购买了完整的套件后，就可以完全应对一般的办公任务了，因此并不需要额外的设备和软件的开销了。但随着互联网的发展，企业对内部联网、协同办公的要求越来越高。当然微软不会对这股强劲的需求坐视不理，从 Office 2003 开始，微软将其命名为“办公系统（Office System）”，旨在为办公环境提供一揽子的解决方案。

Microsoft Office 2003 于 2003 年发布。作为一个整合平台的解决方案，Office System 2003 所包含的产品多得令人瞠目结舌。Office 2003 中文版包括 6 个组件、11 个产品、4 个服务器组件、1 项服务以及解决方案加速软件。Office 2003 中文版产品除了原有的 Office Word、Excel、PowerPoint、OutLook、Access 外，还包括 Office Publisher（发布软件）、Front Page （网页编辑软件）、InfoPath（制表软件）、OneNote（笔记软件）、Visio（工程绘图软件）、Project（项目管理软件）。服务器组件则包括 Office Live Communications Server、SharePointPortal Server、Exchange Server、ProjectServer，以及 Visual Tools 等共计 16 款产品。将这么多的应用整合在一起，直接导致了其昂贵的价格。另外，如果企业要享用 Office 2003 的全部新功能，技术许可费用至少还会增加 10%以上。

Microsoft Office 2007 是为了配合 Windows Vista 而推出的。从这个版本的升级创新，可以看到微软已经充分意识到单机时代的 Office 已经不再是办公软件的发展趋势，沟通、协作将成为今后办公软件的主要特征描述词。从 Office 2007 这个版本开始，微软在 Office 互联网化的历程上投入逐渐加大，其快速发展也让用户对其重拾信心。

从 Office 2010 开始，微软推出了网络免费版本 Office Web Apps，涵盖软件有 Word、Excel、PowerPoint 和 OneNote，用户可利用浏览器来编辑文件和演示文稿等。这个功能因使用体验较差，并未获得用户青睐。

Office 2013 也是一个应用广泛的版本，同时也是目前为止最为臃肿、庞大的版本。

微软办公软件的发展规律，也是传统的工具型软件的一般发展规律。在以卖“拷贝”为主的商业模式中，必须扩展产品功能，丰富品类，来满足更多人的需求，在这一点上微软是成功的。但是对于绝大部分的使用微软办公软件的人来说，譬如 Word（字处理软件），大家日常使用的功能，远远不及其总功能的 1%，但是不得不为 Word 软件的所有功能付费。因为软件是标准化生产出来的，不得不满足所有

人的需要。难以平衡标准化生产和个性化需求之间的矛盾。

另外一个不足之处，就是办公软件产生的文档由用户自行管理，如图 7-7 所示。许多不具备基本计算机知识的人，就被拒之门外。大家不得不通过上培训班，来解决这个问题。事实上，微软也带动了庞大的计算机教育市场。

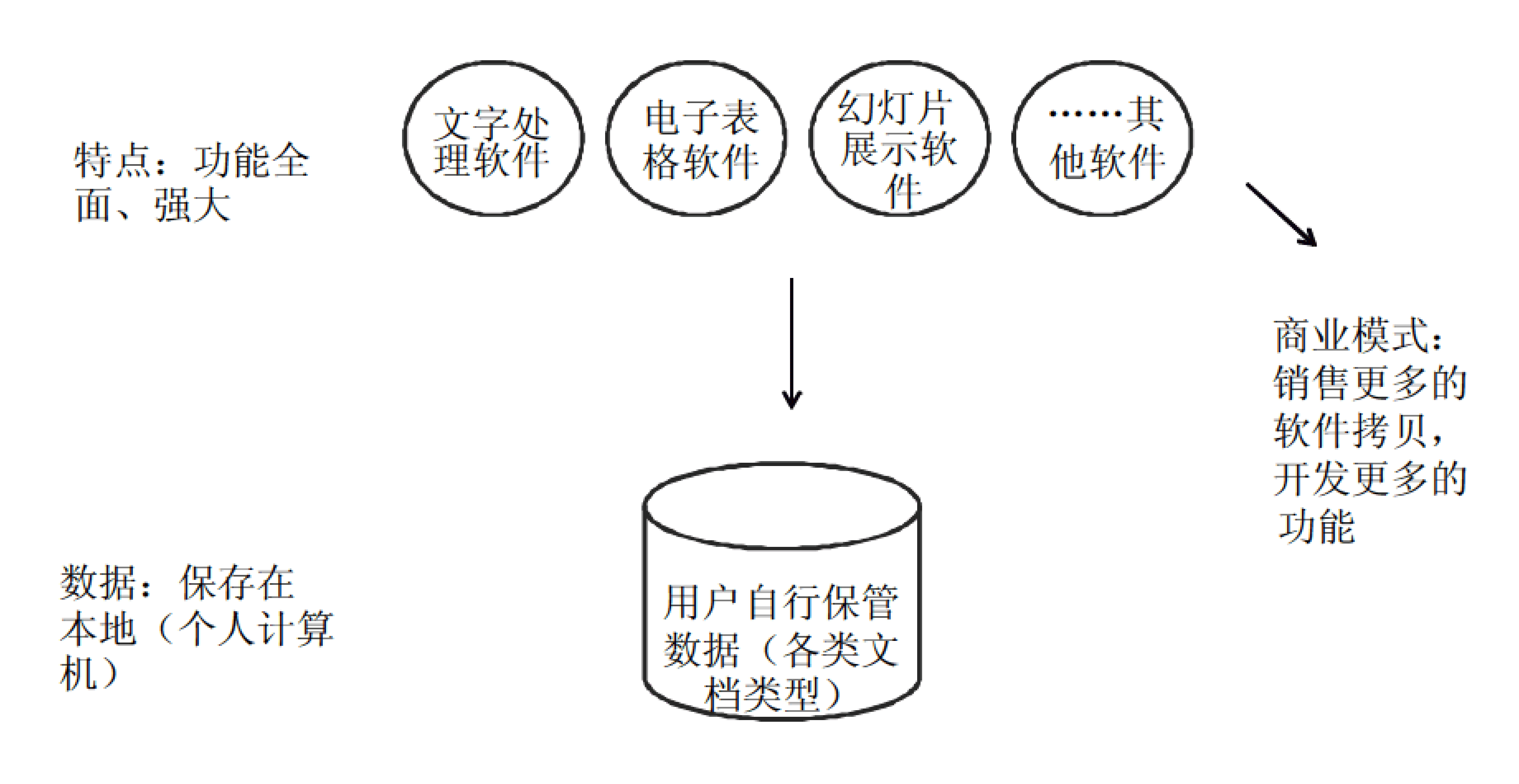


图 7-7　微软办公软件与数据的关系及商业模式

依然用大数据的视角来分析微软办公软件商业模式。微软的盈利来源是销售更多的软件拷贝。在个人计算机主宰的时代，微软办公软件功能扩张路线，市场推广策略行之有效，取得了空前的成功。事实上，这些办公软件产生的大量数据（形形色色存在各种文档之中的数据）蕴含着更加巨大的商业价值。

以前微软的竞争对手，都在办公软件功能上跟微软比拼，同样的商业模式下，它们无法撼动微软利用先发优势建立的垄断地位，逐渐退出这一市场。谷歌公司最先看到办公软件数据的商业前景，建立了数字广告的商业模式，从而开始了和微软的长期竞争之路。

依赖网络浏览器的 Google Docs

谷歌公司是在“不作恶”的口号下发展起来的。所谓“作恶”就是指微软的垄断。微软也把蒸蒸日上的谷歌当作头号竞争对手。谷歌作为后起之秀，叫板微软的

实力，就是来自其崭新的商业模式。谷歌是最早在数据中掘金的公司之一，把数字媒体（数字媒体的详细介绍，参见本书第四章）的商业模式提升到一个崭新的水平。

2006 年，谷歌推出了 Google Docs 服务，包括在线文档、电子表格和演示文稿三类文档，与 Office 办公软件中的 Word、Excel 和 PowerPoint 类似，此工具可以轻松地执行所有基本操作，包括编制项目列表、按列排序、添加表格、添加图像、添加注释、添加公式以及更改字体等。并且，其风格和传统桌面办公处理软件类似，熟悉的风格可以让用户无需学习便轻松上手。

所谓 Web 应用，是指仅仅通过网页浏览器就可以在线使用的一类软件。用户不需要额外下载、安装任何第三方软件。上网即可使用。

具体而言，Google Docs 与微软 Office 有两点最大的不同：Web 应用和免费。

第一，Google Docs 是基于浏览器的一套在线软件，通过谷歌账号登录后便可以使用。与微软 Office 把数据保存到本地的模式不同，Google Docs 将数据保存到云端，用户可以通过浏览器新建、打开、编辑或是删除一个在线文档。同时，同一个工作组中的用户可以共享文档，多用户可以对同一个文档进行实时的编辑和更新，而且这些操作历史也会被保存到云端。协同办公平台创造了许多有趣的使用场景，英国作家 Silvia Hartmann 就利用 Google Docs 玩了个行为艺术，她公开了自己使用 Google Docs 写作的地址，任何人都可以进去看看她的新小说《The Dragon Lords》写到哪里了，如果你碰巧遇到她正在写作，那么可以看到她一个字母一个字母地输入单词直至完成整部小说的过程。

第二，Google Docs 是免费软件。众所周知，微软 Office 主要靠销售更多的软件拷贝盈利。因此为了满足不同类型用户日益差异化的需求，微软必须不断研发新的功能，直接导致的结果是微软 Office 体积越来越大，价格也越来越高。这与互联网的“免费”精神是相悖的。而 Google Docs 完全免费，并且用户可以拥有 5GB 的 Google Drive 存储空间，已经基本满足用户日常使用。对于对存储容量有更高要求的企业用户来说，可选择升级至 25GB 空间，目前其费用为每月 2.49 美元；

还可升级至 100GB 空间，目前每月费用为 4.99 美元；或是升级至 1TB 空间，目前每月收费 49.99 美元。单位存储空间的价格相对于 DropBox 等其他主流云存储平台要低廉一些。同时，由互联网巨头谷歌“生下”的 Google Docs 当然也会“继承”谷歌的优秀商业模式基因，广告也是其主要收入来源之一。《连线》杂志主编克里斯•安德森在《免费》一书中极力推崇“少数人付费，多数人免费或只需花费极少费用享用”的商业模式，他认为免费经济才是商业的未来。

谷歌针对微软，完完全全地反其道而行之，推出一系列的免费 Web 应用。谷歌公司的操作系统是开源的，办公软件是免费的，和微软的商业模式截然相反。谷歌的办公软件产品是以互联网服务的形式提供，用户享受免费服务的同时，自然而然地把文档无偿地交由谷歌来管理和保存。如图 7-8 所示，谷歌利用用户文档中的数据进行加工分析后，可以提供更加精准的广告。

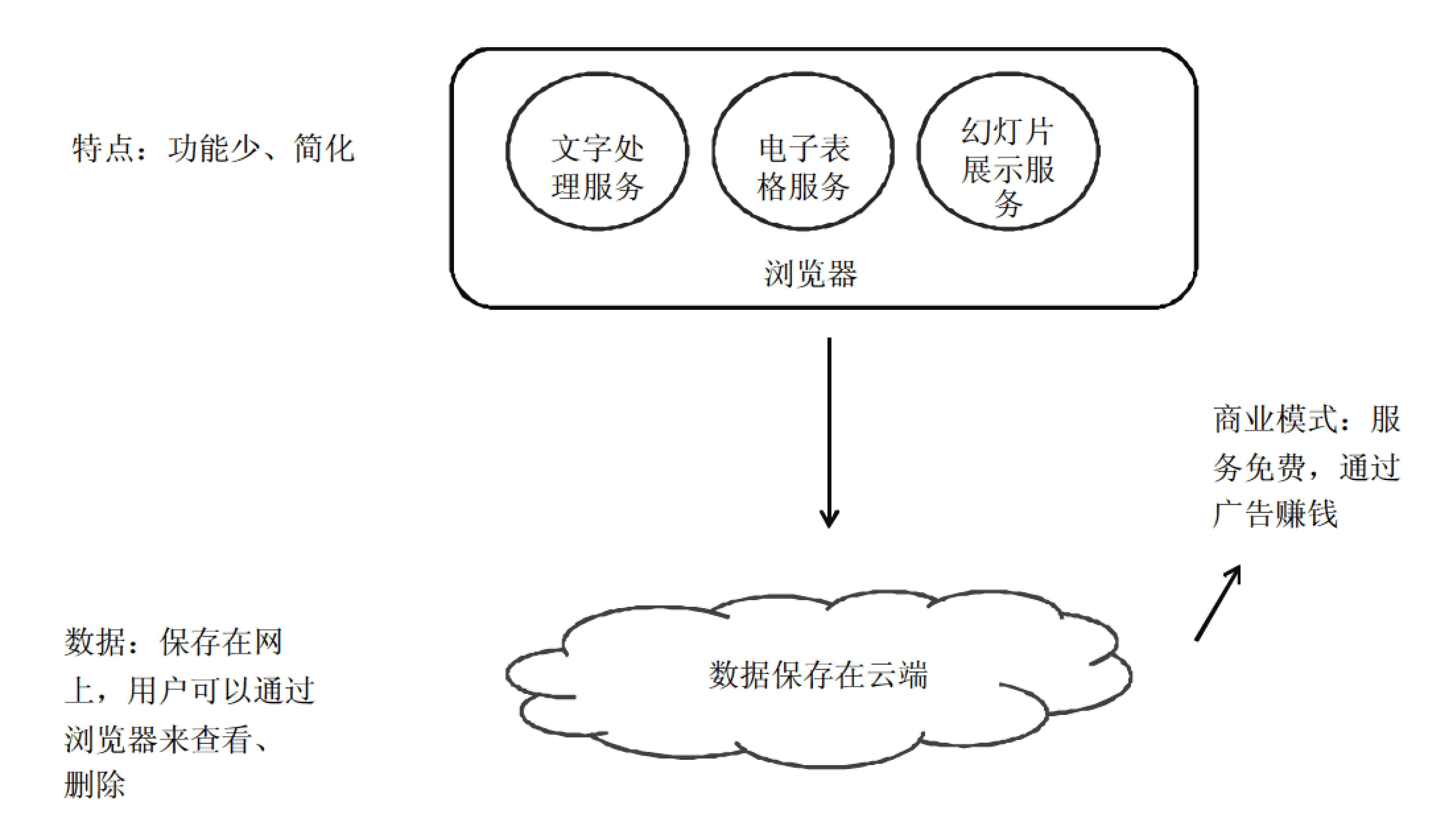


图 7-8　谷歌办公软件与数据的关系及商业模式

谷歌公司并没有公开 Google Docs 的运营数据，人们无从了解它现在的用户数量。这款产品是谷歌早期的开发成果之一，目前并不引人注目。作为一个普通用户，

笔者感觉 Google Docs 在功能方面和微软办公软件相比，距离较大。Google Docs 依然处于不断地发展中，未来的走势有待观察。

Google Docs 和微软 Office 分别处于平衡木的两端，一边是桌面应用的极致，一边是 Web 应用的极致。EverNote 恰恰位于平衡木的中间，取两者之长，避两者之短，这几年飞速成长，引入注目。

泛互联网化的印象笔记（EverNote）

探究起印象笔记的成功原因，有人说是其强大的笔记捕捉功能和先进的文字识别技术，也有人说是其稳定的出色用户体验，但是这些功能以往一些软件都具备。事实上，印象笔记本地文字编辑功能相比微软 Office 办公套件中的笔记软件而言，功能要少很多。但是快速上升的装机量，证明印象笔记是一款广受欢迎的产品。

印象笔记可以在个人计算机、苹果系列电脑以及各种智能手机和平板电脑上使用，如图 7-9 所示。对于一款软件而言，在不同的硬件平台上通用非常重要。这是成为“门户化”特征的基础，也是软件产品和硬件产品在成为“门户”方面的差异。无论用户哪款硬件设备，都可以使用印象笔记。虽然使用的硬件设备不同，但是通过印象笔记却可以保证内容都是一致的、完整的。在个人计算机上，记录的一些文章，在路上掏出手机可以查看；或者参加会议，通过手机录音，回家打开电脑，就能直接编辑、收听这段音频。不需要繁琐的同步操作，印象笔记静悄悄地完成了所有内容在所有设备上的自动同步。这是软件产品门户化的第二个特征：一致的用户体验。

相比之下，早期的微软笔记软件（OneNote）并不具备网络同步功能。升级后的 OneNote 的同步功能仍然显得薄弱很多，在重装软件之后，如果用户需要找回之前备份的数据，需要连接微软的网络存储服务，并且其操作对于普通用户而言非常复杂。在 2012 年之前，OneNote 对移动设备的支持相当乏力，在 iPhone、iPad 和 Android 手机大行其道的当下，对“门户化”的忽视导致了 OneNote 流失了大量用户。

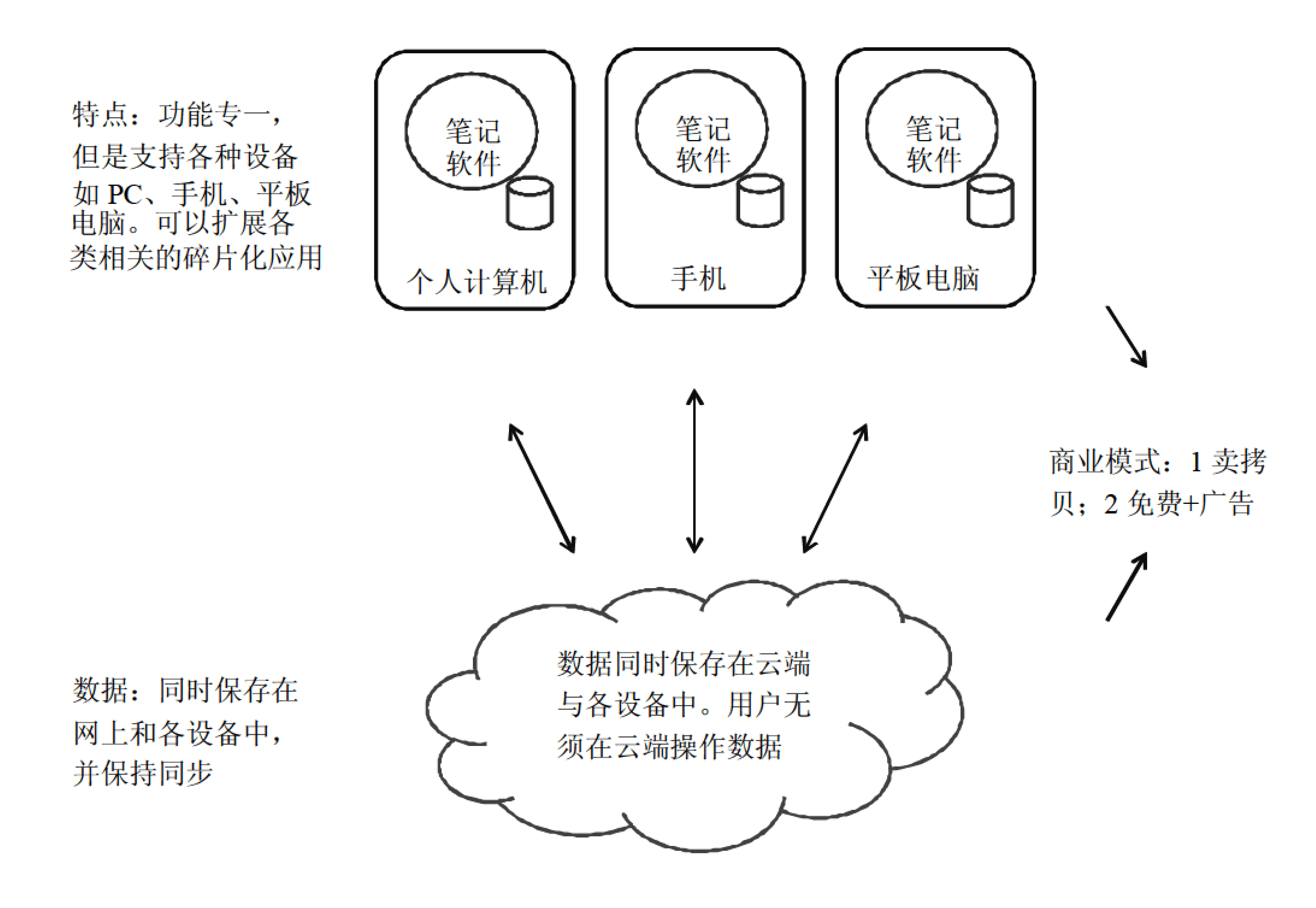


图 7-9　印象笔记与数据的关系及商业模式

印象笔记在抢占“笔记类”应用门户的竞争中拨得头筹后，立即开始了平台化的征途。2010 年，印象笔记推出了自己的应用商店——“百宝箱”，展示了其欲打造与苹果 App Store 分庭抗礼的平台野心。目前其已经收纳了音频、新闻、阅读、生产力、旅行、绘画、手写、无纸化等众多分类的许多应用。截止 2012 年 3 月，百宝箱中的应用已经达数百款，为 EverNote 开发周边应用的第三方开发者也达到了 2 万多人。平台化的意义在于明确了平台的创建者和第三方参与者的利益划分机制，形成众人拾柴火焰高的局面，如图 7-10 所示。

此处通过印象笔记百宝箱中的几款“小”应用，来阐述“碎片化”的意义和要点。

前文笔者反复强调过一点，就是碎片化的机制解决了标准化生产和个性化需求之间的矛盾。微软办公软件为满足所有用户的需求，不得不把 Word（字处理软件）、excel（电子表格软件）变得庞大无比，而大多数用户仅仅使用其中不到 1%的功能。

这种方式依然是工业时代标准化生产的思维，既增加了普通用户的学习成本，也让大家花费了不必要的金钱。当然，盗版是另外一回事情。

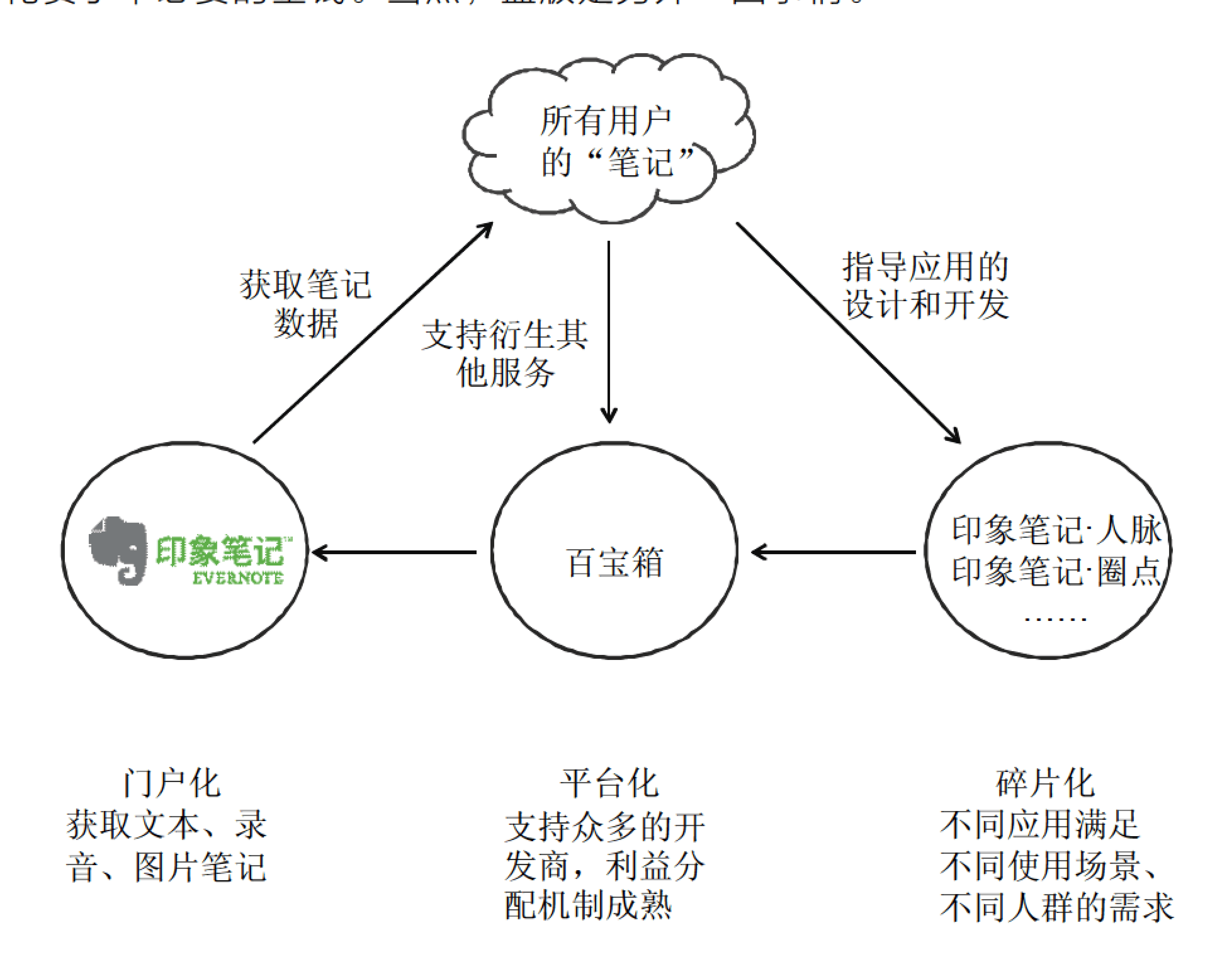


图 7-10　印象笔记的泛互联范式

印象笔记就聪明得多。它仅仅实现了满足 80%用户需求的 20%的功能，其余 80%的功能通过其他的应用来实现。譬如手写功能，保存原始的笔迹，对一部分用户而言有相当的吸引力，尤其是当用户写一笔好书法的时候，但是在个人计算机上，手写输入不一定比得过键盘输入的速度，而在 iPad 等平板电脑上，连续的手写识别，肯定要远远强过平板电脑自带的手写输入功能。总之，手写功能是需要的，但是小众的，只适合某些特定的场合。

沿袭工业时代标准化生产思维，就不得不为印象笔记添加手写功能模块。不管用户是否需要，软件中已经包含了这个功能。但“碎片化”思想完全不同，“两个独立，一个融合”。首先手写功能作为一款独立的应用存在，其次用户需要单独付费购

买，但是使用过程中产生的数据，却是和印象笔记中的“笔记”融合在一起。也就是说，当你在开会时，在 iPad 上手写记录的会议纲要，打开个人计算机上的印象笔记软件，就能立刻查看和修改。

具备程序设计素养的读者，一定会明白，碎片化对应用程序功能的规划和设计要求是非常高的，把“高内聚，低耦合”的设计思想离散化，在互联网之上，保障应用程序数据的一致性。笔者曾经一篇谈“企业架构”的文章中，说要高度重视“数据架构”的思想。在大数据时代，必须把“数据架构”思维扩张到整合互联网之上，这是高级架构师们必须面对的课题。继续探讨数据融合的原则和实现路径，远远超出本书的主旨，就留给架构师伤脑筋吧。

在商业模式上，碎片化的应用也有大幅突破，完美地体现了《免费》、《长尾》两本书中提到的主旨。通过免费的方式提供满足 80%用户的功能，大家不需要支付任何费用，就可以自由享用印象笔记软件，而且完全可以满足绝大多数用户的需要。这里需要着重指出的是，免费不意味着缺斤短两，也不代表服务质量低劣。具体提供哪些免费功能，是要精心规划的，但原则是吸引尽量广泛的用户。20%用户特殊的需要，通过大量的碎片化应用来满足，也就是长尾。这部分应用大部分是收费的。泛互联网化模式中，收费也和原有的软件定价有了本质的差别。价格非常便宜，也就是“微支付”。微支付结合碎片化的长尾类应用，迸发出惊人的商业力量。

对比印象笔记、Google Docs、微软办公软件，可以清晰地发现泛互联网化软件商业的前景，如图 7-11 所示。毫无疑问，印象笔记目前最受资本市场追捧和青睐。前面介绍的苹果和印象笔记，都是针对个人的消费类产品。在一向保守的企业市场中是否也有类似的应用呢？又能带来哪些改变呢？

当笔者完成这节的写作，闭目深思。人们长期被禁锢在以“文件”为中心的思维，所有的软件都是产生各种各样的文件。文件的结构和复杂性，增加了人们获取文件中蕴含的数据和信息的障碍。文件这个概念，或许应该被丢弃。数据和信息将以崭新的面貌被重组和分享，也许人们已经临近另一场信息革命的边缘。

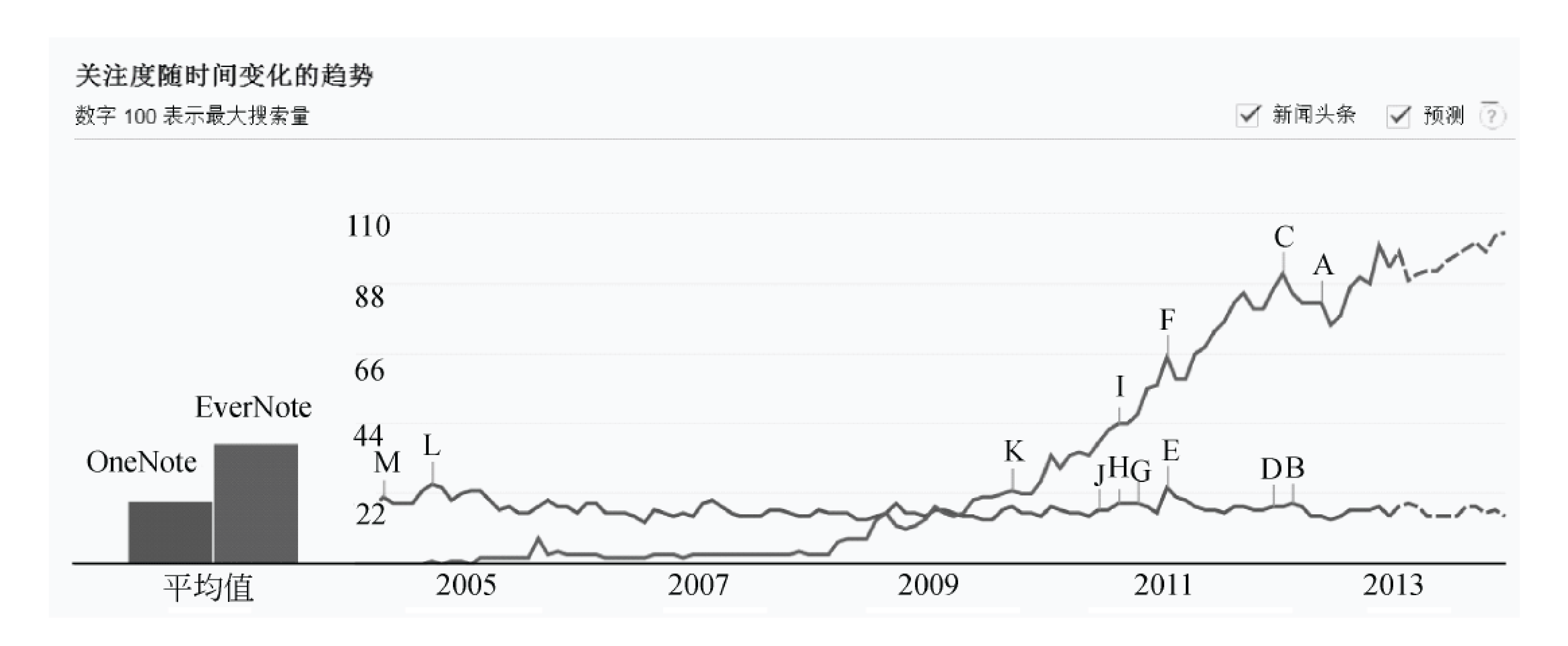


图 7-11　谷歌趋势中 OneNote 和 EverNote 搜索次数对比

第三节　旺铺助手——小软件的大梦想

提要：

1. 泛互联范式在企业市场的应用，是我国企业信息化市场寻寻觅觅十数年，探索出来的一条行之有效的路径。
2. 纯粹的基于浏览器的企业应用，即所谓的 SaaS 模式，在泛互联范式的强烈反差下必须重新思考未来。必须重视终端，不能仅仅依赖浏览器；必须重视数据，从数据中淘金是上上良策。
3. 美国的 Salesforce 成功，在中国不具备代表性。中国的商业环境、信息基础设施建设水平、企业人才结构等，更支持泛互联范式落地生根。

在路演推介用友软件的过程中，笔者为许多投资人讲了旺铺助手这款小软件的例子，引起很多人的兴趣。当然对于用友庞大的体量而言，旺铺助手短期内不可能

大幅增加用友的营业收入，但是从中的确可以发现，商业领域泛互联网化软件的广阔空间。顺便强调一下，本书中提到的案例，如果涉及上市公司，并不代表笔者的评级立场，只是为了更让大家更容易理解一些产业现象。

中小企业信息化——寻寻觅觅的十年后，依然在灯火阑珊处

中小企业以其庞大的数量，一直令大型公司垂涎。据统计，2011 年全国有 4600 多万家中小型企业，其中仅仅不足 300 万家获得有支持的信息服务。微软进入企业应用软件市场后，力图占领更多的小型企业，但是收效甚微。从 2000 年以来，笔者亲身经历了三次中小企业管理信息化浪潮，前两次皆无功而返，如图 7-12 所示。

2000 年前后，正值.COM 泡沫顶峰，ISP[①]、ICP[②]、ASP[③]等等概念层出不穷。尤以 ASP（应用托管服务提供商）概念因为有微软、IBM、Sun 等公司力推，而领一时之风骚。所谓的应用托管服务，就是把软件安装在远程的服务器，用户通过互联网来使用。允许小型企业客户多人使用一套软件的授权许可，这种做法虽然降低了用户的采购成本，但是在使用中更加繁琐。大家无法理解，好端端的 Word 软件非要在网上用。加上当时带宽严重不足、应用托管服务商要和盗版作斗争等原因，ASP 最终毫无悬念地以失败告终。这完全是软件提供商一厢情愿的做法。

2002 年以后，微软、IBM 等公司又重提 SMB（即中小型企业）的概念，希望把给大型企业提供的业务管理软件，经过简化后，卖给中小企业，联合许多合作伙

① ISP：互联网服务提供商（Internet Service Provider，ISP，又称因特网服务提供者、互联网服务供应商、互联网服务提供者）即指提供互联网服务的公司。通常大型的电信公司都会兼任互联网供应商。

② ICP：线上内容提供者（Internet Content Provider，ICP，又译互联网内容供应商），其业务范围为向用户提供互联网信息服务和增值业务，主要提供有智慧财产权的数字内容产品与娱乐，包括期刊、杂志、新闻、CD、On-line Game 等。线上内容提供者模式的收益包括广告收入、下载收入、订阅收入、中介佣金收入等。但 ICP 目前受到消费者自行创造内容的 Web 2.0 的强大威胁。

③ ASP：应用服务提供商（Application Service Provider，ASP），是一种服务的供应商，该服务名为电脑应用软件，尤如租车公司，令顾客有更多的选择，共享不菲的软件。

伴共同开拓 SMB 市场。这个出发点也不错，国际巨头们提供标准化的、简化的解决方案，众多合作伙伴们帮着中小企业削足适履。事实上，中国的中小企业发展变化非常迅速，可能三五年就成长为中大型公司，也可能两三年就销声匿迹。希望用标准化的方法为中小企业提供业务管理软件，无疑是死路一条。

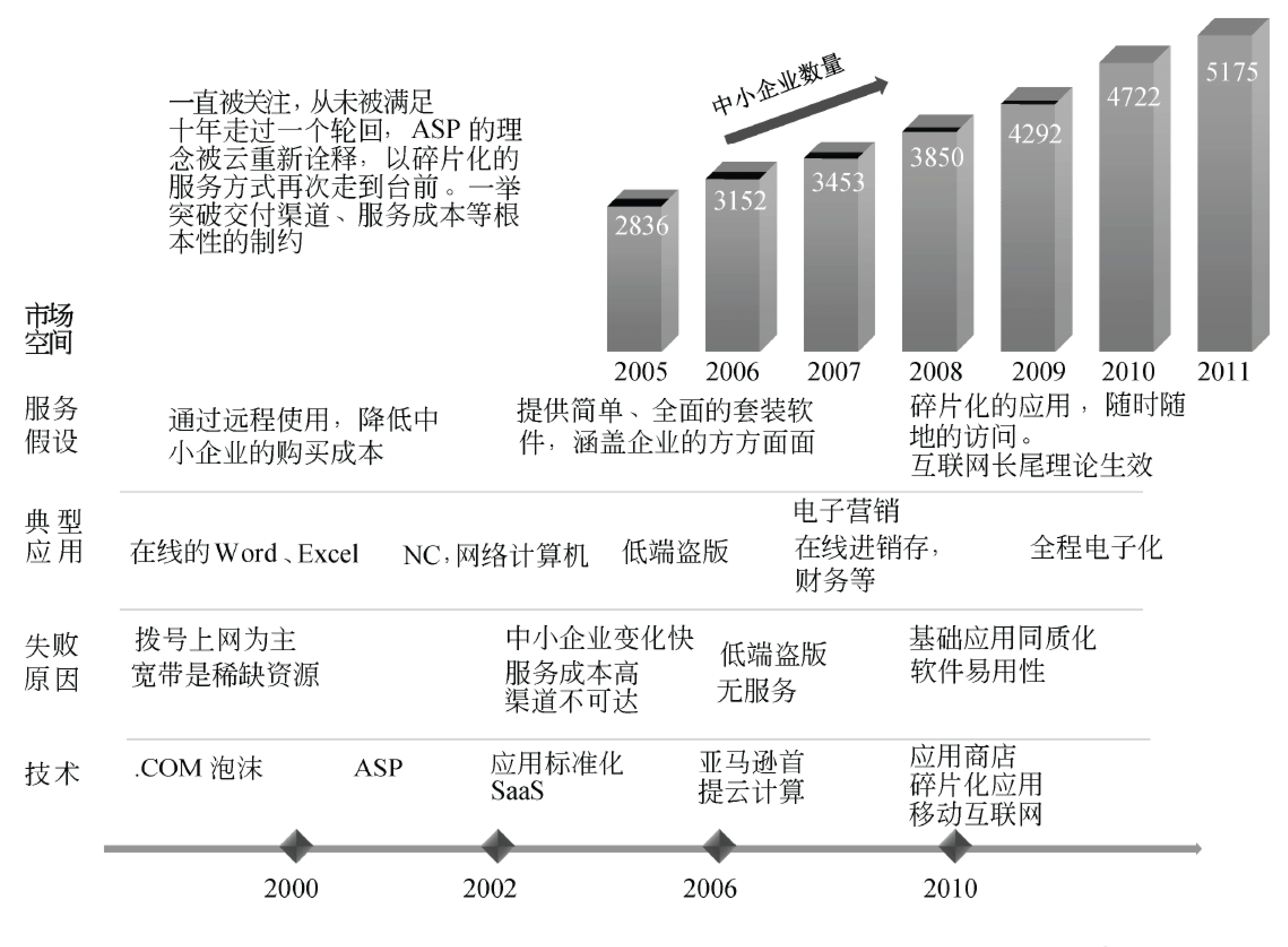


图 7-12　中小企业信息化，寻寻觅觅十年后，依然在灯火阑珊处[①]

2006 年，云计算模式的出现给中小企业管理服务带来了一丝曙光，一时间中小企业信息化柳暗花明。通过云的形式，提供中小企业所必须的服务。这一创新的模式经过美国一家公司 Salesforce 验证成功，SaaS[②]（软件即服务）的理念深入人

① 参见国金证券个股深度研究报告《SaaS 一小步，用友一大步——云的梦想》，第 3 页。

② SaaS：软件即服务（Software as a Service，SaaS）有时被作为“即需即用软件”（on-demand software，即“一经要求，即可使用”）提及。它是一种软件交付模式，在这种交付模式中，软件及其相关的数据在云端集中式地托管。用户通常使用瘦客户端，通过一个万维网浏览器来访问软件即服务。

心。的确 Salesforce 不再销售软件，而是通过互联网出租“服务”。中小企业只要每月付很少的租金就可以获得完整的软件使用权，也不用担心软件安装、数据管理等等问题。

下面先详细介绍在美国获得成功的 Salesforce 的特点。

Salesforce 简介

2004 年 6 月，Salesforce 公司在纽约证券交易所成功上市。2004 年其收入达到 1.75 亿美元，2012 年突破 20 亿美元，达到 22.6 亿美元。Salesforce 的创始人贝尼奥・马克是一个具有传奇色彩的人物，他创立 Salesforce 之前是 Oracle 高级副总裁，当时才 27 岁，是 Oracle 历史上最年轻的高级副总裁。20 世纪末，互联网的发展出现了一个高潮。当时，见尼奥・马克预见到，随着互联网的发展和宽带的普及，会有越来越多的企业通过互联网得到一些软件的服务。于是他在 1999 年成立了 Salesforce 公司，开始对 SaaS 业务模式进行探索。

Salesforce 通过云计算的业务模式，解决了用户购买硬件、开发软件等前期投资以及复杂的后台管理问题等麻烦。公司强大的在线开发平台，允许用户与独立软件供应商定制并整合其产品，同时建立他们各自所需的应用软件，解决了标准化产品和个性化服务的问题。事实上，Salesforce 把个性化需求的问题抛给了用户，公司仅仅提供技术支持。

Salesforce 进入中国，最早是和神州数码成立合资公司共同开拓国内的市场，做了几年没有起色。尽管 Salesforce 功能很强大，很灵活，但是国内的中小企业根本就没有会使用这个工具的人。说白了，客户自己定制虽然比开发的难度要小一点，还是远远超出了绝大部分中小企业员工的计算机水平。况且中小企业发展迅速，随意的业务变更更是让 IT 支持人员焦头烂额。

Saleforce 期待让中国用户自己解决个性化的需求？天方夜谭。

中国 Salesforce 模仿秀步履艰难

中国公司看到 Salesforce 在纳斯达克风光上市，业务蒸蒸日上，当然是坐不住了，一拥而上，刹那间 SaaS 概念传遍大江南北，Salesforce 成为中国创业者当中新的灯塔。可惜，站在 2012 年年尾，回顾那些模仿秀们当初的豪言壮语，不禁唏嘘不已。8 年过去，几乎很少有同类公司的营业收入超过 1 亿元人民币的。

究其原因，这些 Salesforce 模仿秀是在用互联网提供标准化的产品，并没有解决用户个性化的需求问题。Salesforce 在美国，通过有限的技术支持，可以把个性化需求甩给用户 DIY。但是在中国，此路不通。

几个月前，笔者曾经和一个“模仿秀”公司的董事长争论：“企业 SaaS 基础服务是否应该免费”。他的答案是不能。理由是企业担心免费的服务靠不住。“你要是免费，人家反而不敢用了。”这话听起来不无道理，但是错的。

《免费》和《长尾》书中，分别提出的免费理论和长尾理论[①]是互联网通行的法则之一。注意，这里的免费指的是接近于零的极低价格。长尾理论是对二八定律的颠覆，公司的利润不再依赖传统的 20%的“优质客户”，而是许许多多原先被忽视的客户，他们数量庞大，足以让你挣得盆满钵满。从公司产品的角度分析，拳头产品主打市场的老套路将趋末路，而免费则是吸引庞大客户群的杀手锏。免费不是给人家提供不可靠的服务，而是把最重要、最常用、用户最广泛的应用，做到极致的简化、易用和稳定后，再用免费的模式培育广泛的客户群，用长尾效应来盈利。

所以 SaaS 基础服务是否免费的问题，归根到底是能否找到一款可以带来长尾效应的终端（注意，这里的终端可以是软件，也可以是硬件）。如果你不能挖掘此类

---

① 长尾（The Long Tail），或译长尾效应，最初由《连线》的总编辑克里斯·安德森（Chris Anderson）于 2004 年发表在自家的杂志中，用来描述诸如亚马逊公司、Netflix 和 Real.com/Rhapsody 之类的网站的商业和经济模式。它是指那些原来不受到重视的销量小但种类多的产品或服务由于总量巨大，累积起来的总收益超过主流产品的现象。在互联网领域，长尾效应尤为显著。

终端，根本就不能谈免费的问题。这类终端一定符合几个特点：第一，使用的人群要足够多；第二，使用的频率要足够高；第三，使用非常便利，具有门户化特征。再者，必须要化解标准化软件和个性化需求之间的矛盾，有效地控制成本并提升客户体验。

在企业服务市场，笔者的确发现了这样的一款终端——旺铺助手。

旺铺助手

旺铺助手是中国电信与用友合作推出的一款通信增值服务，主要适用于客户群比较固定以及电话、短信等业务使用频繁的中小企业和普通聚类商户，如小型批发企业、具有配送性质的商铺、街头门店、初创阶段的小企业，或者服务对象相对稳定的物业公司等。其主要功能包括记录客户资料和订单、接收短信、来电提醒等等，基础版本费为每月 10 元。

中国电信官方给出的宣传资料上称："旺铺助手其核心是将中国电信综合通信能力与企业内部业务流程相结合，在常用的客户管理以及进销存应用中嵌入短信收发、拨号呼叫等通信功能，是典型的信息化融合型产品。旺铺助手是中国电信通过智能化终端，融合 C 网、固网等通信能力，以客户管理以及进销存应用为主要切入点，为中小企业和普通聚类用户提供的集客户和订单管理、商品管理、来电弹屏、点击拨号、短信收发等功能于一体的综合应用服务。旺铺助手可帮助中小企业有效聚拢客户并开展针对性促销，是中国电信关注中小企业信息化需求、降低信息化建设成本而推出的业务。"

看起来平淡无奇，这么个小东西，最多也就是玩个跨界呗，融合简单的资料管理和通话管理功能，有什么大用呢？事实上，这款软件非常受欢迎，自 2011 年 5 月推出后，短短时间内就发展了 20 多万用户。

现实中的使用场景是比较有趣的，当有人打电话进来时，软件会提示来电，包括来电姓名、地区等信息；如果公司同时购买了客户关系管理模块，提示中就会增加更多的提示信息，如公司和其业务往来资料；如果公司又购买了财务管理模块，

旺铺助手甚至可以在你接电话之前，就会提示公司在与此人的合作过程中，赚了多少钱，或者是否还有应收账款等等信息。每个功能模块的月服务费非常便宜，几元钱到十几元钱不等。

小微企业与客户沟通，使用最多的工具依然是电话，无论是移动电话还是固定电话。旺铺助手在普通的资料管理功能上，和频繁的通话融合在一起。每次通话，软件都会自动记录来电详询情况，都会弹出窗口提示这款软件的存在，事实上形成了公司使用的管理软件入口和门户。上面提到的客户关系管理模块、财务管理模块等等，都是以“小应用”的形式供客户任意下载的，即下即用。旺铺助手又同时具备了软件推广和获取服务渠道的特征。泛互联网化应用与系统桌面应用、Web 应用的对比见表 7-1。

表 7-1　泛互联网化应用与传统桌面应用、Web 应用的对比

	传统桌面应用	Web 应用	泛互联网化应用
销售方式	卖软件“拷贝”	租用服务	卖拷贝+租用服务
数据	客户本地	宿主网络服务器	客户本地+宿主网络服务器
工作方式	缺少网络功能	只能在线使用	基本功能离线完成；在线获取更佳客户体验，相关资讯随需可得
门户化特征	无	有	有
平台化特征	无	有	有
碎片化特征	无	无	有

旺铺助手类的泛互联范式的成功，预示产业界终于摸索出了一条服务于企业客户的新路。这条道路的成功依然受苹果“应用商店”模式的启发，但是不同于对 Salesforce 的跟风式模仿，而是结合小微企业的需求特征，并充分结合了传统桌面应用和 Web 应用的优点，成为了发挥互联网潜力的一种新模式，如图 7-13 所示。

这种模式既不同于完全依赖移动互联设备的做法，又区别于基于浏览器的 Web 应用，完全是从客户迫切需要解决的实际需求出发，拥有强大的生命力。正是因为此，笔者称之为泛互联范式，并详加解释。

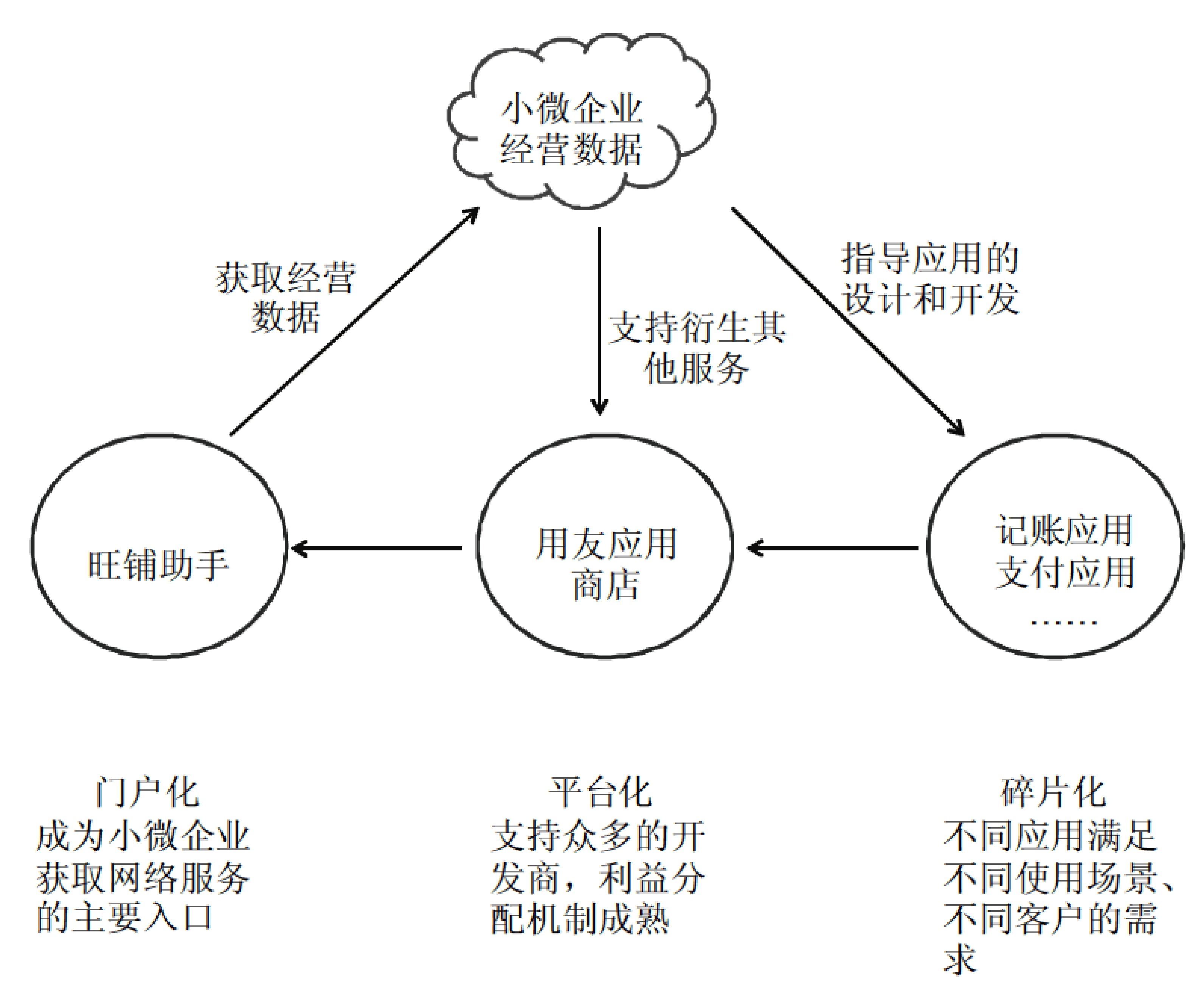


图 7-13　旺铺助手带动的“泛互联范式”

第四节　泛互联范式启动大数据飞轮效应

有些公司在经营中，已经积累了大量的数据，如互联网公司、电信运营商、银行等。这些公司面临的首要问题是发展数据挖掘技术，充分释放数据资产中蕴含的商业价值。

但对于初创的公司，或者新的业务方向，泛互联范式提供了一个可能的路径来启动大数据的飞轮。一点一滴地收集数据的过程是艰辛和枯燥的，必须精心地打磨终端，让更多的人接受、使用、喜爱它。有了终端基础，才有积累数据的可能。围绕这款终端产品，构建起利益分享的机制，吸引更多的人、更多的团队开发相关的碎片化应用，收集更多的数据。在这种日

飞轮效应指为了使静止的巨大的飞轮转动起来，一开始你必须使很大的力气，一圈一圈反复地推，每转一圈都很费力，但是每一圈的努力都不会白费，飞轮会转动得越来越快。达到某一临界点后，飞轮的重力和冲力会成为推动力的一部分。这时，你无须再费更大的力气，飞轮依旧会快速转动，而且不停地转动。这就是“飞轮效应”。

——引自百科全书

复一日的艰苦努力下，大数据的飞轮开始慢慢转动。当数据的积累足够多时，就可以利用这些精心积累的数据，展开衍生的商业模式，如图 7-14 所示。

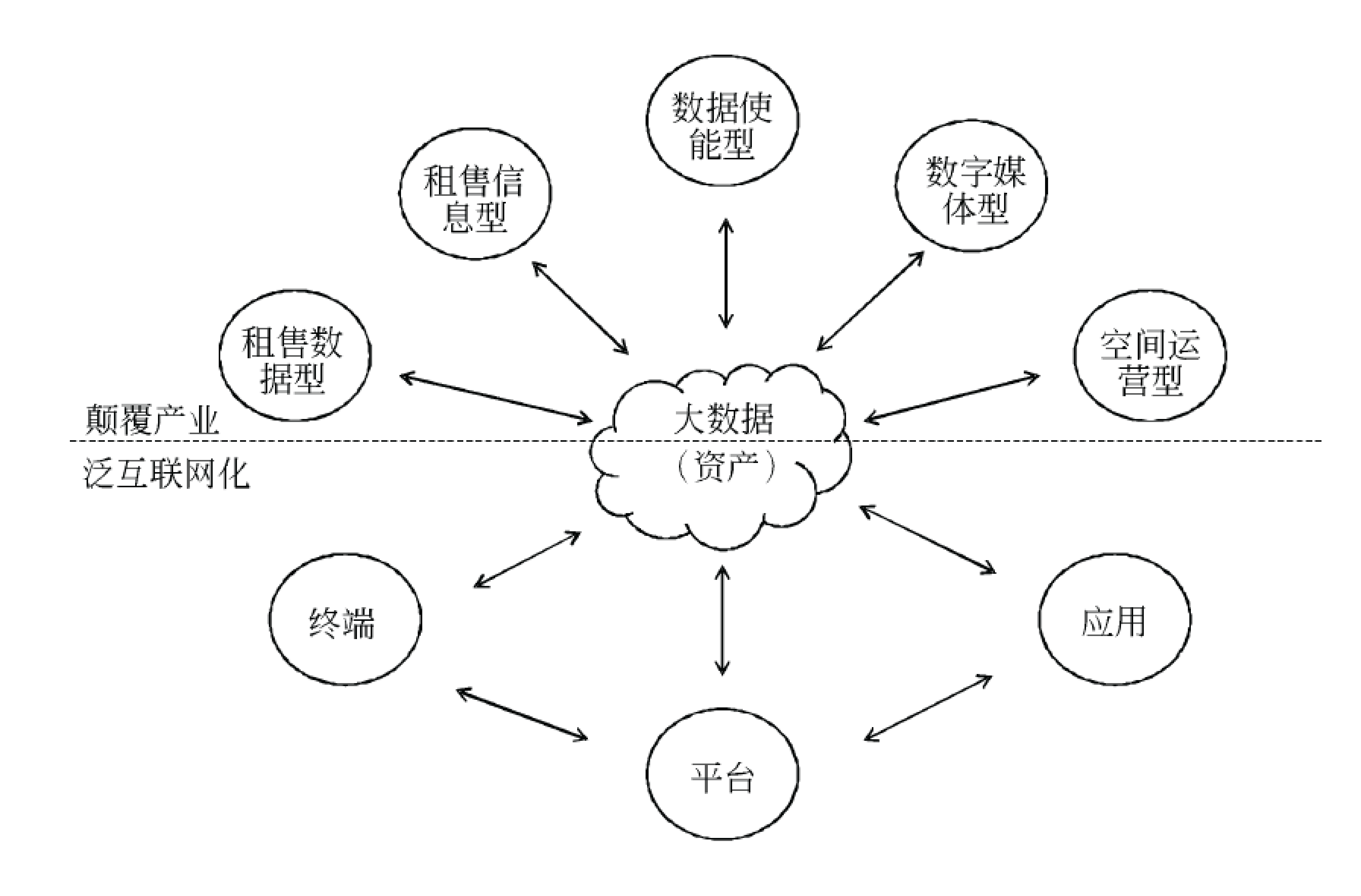


图 7-14　大数据飞轮效应

在和产业界频繁接触的过程中，笔者接触到许多有思想、有志向的人，谈起互联网领域的创业，都是神采飞扬，但是真正理解泛互联范式的人并不多见。他们往往忽视了终端的重要性，或者过于依赖终端而忽视了终端带动的平台、应用、大数据。在推动大数据飞轮转动的时候，终端是至关重要的抓手，平台是聚拢合作伙伴的机制，而应用则满足用户多方面的需求。比拼到最后，决胜还是要靠对数据资产的垄断和运用。

苹果这个模范生，是以 iPod、iPhone、iPad 等一系列创新型的硬件终端启动了其大数据飞轮。腾讯是中国互联网公司的代表，凭借其无处不在的 QQ 聊天软件，收集大量的用户行为信息，也是大数据的典范。其 2012 年初推出的 QQ 圈子，甚至能帮助大家找到“曾经上铺的同学”。亚马逊拥有全球最多的商品交易数据，它是凭借电子商务应用，推动了大数据的飞轮。

苹果和谷歌公司由合作到反目的过程，以及观察它们在大数据飞轮上的战略定位，非常有趣，也有助于大家理解大数据的意义。

苹果与谷歌的数据之争

最近苹果和谷歌之间的争斗颇为引人注目。这两家公司曾经有一段蜜月时光，谷歌的总裁施密特一度担任苹果公司的独立董事。谷歌是旗帜鲜明的互联网公司，提供免费的互联网服务；苹果是特立独行的设备制造商，生产让人过目不忘的电脑、音乐播放器和智能手机。二者的收入来源也完全不同，谷歌的主要收入来源是广告收入，占其全部营业收入的 90%以上，而苹果销售硬件的收入几乎可以使得互联网服务的收入忽略不计。看起来两家是井水不犯河水。

但是这两家巨头发展的最终模式，就是在泛互联范式中拥有完整的布局。谷歌目前的短板是终端，苹果的短板是变现数据资产的能力。如果谷歌不去占领终端市场，则有可能被苹果扼住其收集数据的咽喉要道，从而使谷歌的数据面临枯竭的危险；如果苹果不去开发关键应用，则辛辛苦苦建立的商业帝国，将成为源源不断为谷歌输送数据的通道。所以这两家的战争在终端层面开始，但是竞争的核心是用户的数据。

苹果公司门户化的终端产品非常强势，iPhone 赚取了大量的利润，iPad 几乎垄断了平板电脑产业。在苹果的收入结构中，iPhone 和 iPad 贡献了绝大部分，因此它可以更多的向第三方，那些碎片化应用的开发者们让利。但是苹果却不能允许自家平台上用户数据的流失，其推出的 iCloud、Game Center 都是直接获取用户数据的方式。

谷歌公司一直是从大数据里面淘金的。门户化的终端产品能盈利最好，不能盈利，只要不亏损也行，它要的是能够搜集用户行为数据的碎片化应用可以遍布各种

硬件终端。所以，谷歌的 Android 操作系统是免费授予手机开发商使用的。免费使用 Android 的前提条件之一：必须包括谷歌的搜索、地图、邮件等服务。最近谷歌刚刚推出 4 核的智能手机，价格不足 2000 元。谷歌的战略很明确，我要“大数据”，其他都可以商量。

苹果公司和谷歌曾是紧密合作的伙伴，它们曾经同心协力对付它们共同的敌人——微软。但是最终对于“大数据”的争夺，使这对盟友反目。

苹果推出 iPhone 后，直到最后一刻，才决定在首页上增加谷歌“地图”应用。没想到地图应用如此受欢迎，大量的第三方提供的应用中，都使用了谷歌地图。这意味着，苹果通过昂贵的 iPhone 智能手机，吸引来的用户群、访问流量，几乎都被谷歌拿走了。如果放任这个局面继续下去，苹果很可能沦为谷歌公司收集用户行为数据的一个渠道。而未来是大数据的竞争，失去了数据的控制权，苹果无疑会失去未来。

终端强势的苹果开始限制应用强势的谷歌，而谷歌也觉察到苹果的企图。于是大家看到现在的这幅商业图景：谷歌推免费的智能手机操作系统，推低价易用的智能手机；苹果开始提供自家免费的地图服务，剔除手机中谷歌的产品。无疑，这两家都把数据当成最为重要的资产来看待。

对于汽车互联网的预测

2012 年 7 月 14 日，通用汽车公司解除了与惠普的 IT 服务合同。而此前惠普为通用汽车服务已经长达 25 年，惠普承担了通用汽车 90%的信息服务。通用汽车要在未来三到五年招聘 1 万名 IT 员工，成立 IT 创新中心，对内部信息化、汽车设计、制造流程、流水线的平台、供应链、销售渠道、质量控制等 IT 服务进行内包。

事实上，汽车业内的有识之士，也将汽车看成另类的“移动终端”，开始全面拥抱互联网，就像苹果的 iPhone 一样，承载巨量的信息服务。

用泛互联范式来思考，汽车无非是另一个移动终端。当人们坐在自动驾驶的汽车中，可以使用的车载互联网应用，不会仅仅局限于导航服务，而是将变得非常丰富，也许和驾驶另外一辆车的同伴闲聊两句，再让汽车帮忙找个餐馆。

相比于消费电子产品的更新周期，汽车的更新换代无疑慢如蜗牛。汽车互联网领域，也并非汽车制造商的禁脔之地。苹果、谷歌等都是站在汽车业门口的“野蛮人”。笔者个人认为谷歌的能量更为巨大，因为在谷歌看来，汽车是另外一个数据来源的渠道。谷歌牢牢控制大数据门户的策略，很可能在汽车市场上演。

预测 HTML5 和原生 App 创业者的投资价值

对投资者而言，可能对业界广泛争论的 HTML5 技术（基于浏览器，理论上任何浏览器都可以通用，因而具备了跨平台的特征）和原生 App（针对某款移动设备，开发的应用。如果要移植到其他平台，需要重新开发）难以取舍。对分析师而言，也需要判断这两种方式走向的前景。

利用泛互联范式，可以简单的预测。HTML5 技术依然是标准化的思维，似乎并没有为个性化服务提供解决之道。原生 App 不管采用哪些开发技术，目标是简单明确的，满足用户特定需求，跨平台的事情与用户无关。也许很长一段时间内，尽管采用 HTML5 的开发技术，也需要披着原生 App 的皮。

这个结论看起来比较武断，毕竟没有仔细去研究 HTML5 的技术细节，也没有去广泛的调研，只是站在用户的角度来思考和预测。这个结论正确与否，尚待检验。

表 7-2 罗列了一些符合泛互联网化特征的软件产品，供读者参考。

表 7-2　部分符合泛互联范式的产品

产品	门户特征	人群特征	主要盈利来源	关联碎片化应用
谷歌搜索	搜索	主动搜寻资料的用户	广告	邮件、日历、地图、企业搜索、客户关系管理，Google+
360 安全卫士	安全	对电脑安全知之甚少的用户	广告 卖流量 第三方应用推广	防火墙、网盾、网购保镖、极速浏览器、系统急救箱
产品	门户特征	人群特征	主要盈利来源	关联碎片化应用
QQ	即时通信	强社会联系的用户	增值服务 交叉推广 广告	音乐、游戏、支付、旅行等方方面面的网上生活
新浪微博	社交	弱社会联系的用户	广告 卖流量 增值服务	微盘、微访问、游戏、微音乐、微电台
雅昌艺术	垂直门户	艺术品鉴赏、收藏	广告、中介、拍卖、活动	SNS、画作推广
旺铺助手	垂直门户	微型企业	服务费，交叉推广	财务管理、业务管理、客户管理
搜狗输入法	通用	所有人群	产品交叉推广	搜狗浏览器等
印象笔记	办公软件	所有人群	服务费 广告费	一些相关应用，如 Web clip 等

导读：

1．大数据在微观层面赋予人们高效了解个人、单体的能力，历史上没有任何企业能够像互联网平台型公司那样，用新的技术、新的模式，无限靠近消费者，并大规模地收集消费者的数据。它们对未来的预判能力，是传统企业无法企及，甚至难以想象的。

2．总体来看，大数据在三个方面深刻影响企业文化、战略和组织结构。第一，大数据将颠覆传统的价值链，使工业时代以生产、采购中心的模式，向信息时代以消费者为中心的模式转变。第二，数据驱动的产业链合作，使社会化协作、网络化生产成为现实。第三，大数据使企业的疆界变得模糊，员工和消费者的界限逐渐消弭；使企业的组织结构发生倒置，企业文化和战略随之调整，需要适应大数据时代发展的要求。

第八章

大数据掀起的企业组织变革

大数据首先是一种思维方式，必须融入到企业的每一个毛细血管中。

——笔者

德鲁克认为企业存在的理由就是创造客户，企业家必须要清楚“谁是你的客户”[1]。“以客户为中心”的企业运用大数据，创造出一个个精彩的商业模式：多变平台、定制化生产、客户参与产品创新……这些商业模式都有一个共同的主线，就是它们内部的价值链以客户需求为起点，而不是传统地以企业的资产或核心能力为起点。

运用大数据，客户细分从大众化、模块化向颗粒化、个人化方向转变，客户定义从标准化向个性化转变，需求信息从阶段性梳理向实时性处理转变，客户实质性介入并掌握了企业的控制权，“以客户为中心”成为企业运营的核心主线，驱动组织内部价值链向智能化和柔性化方向发展，如图 8-1 所示。

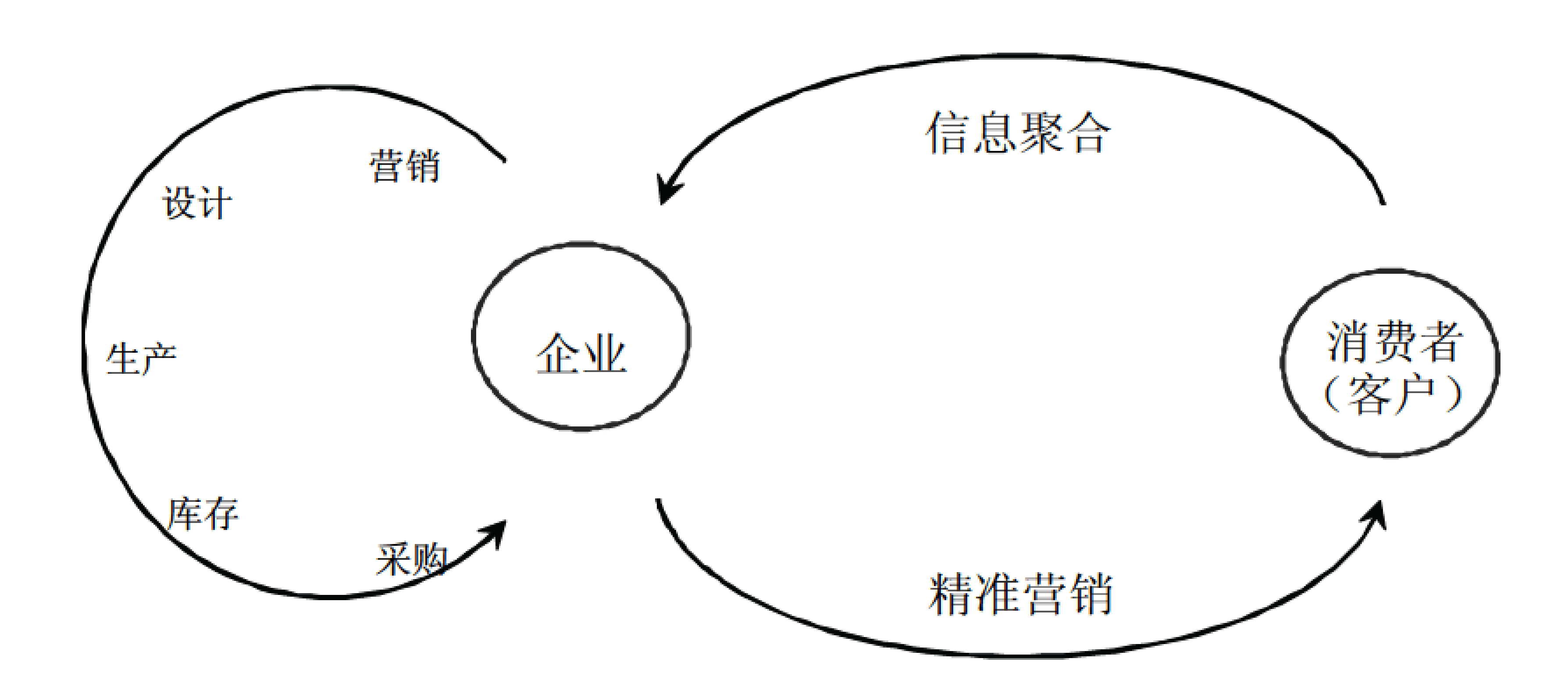


图 8-1 以消费者为中心，进行组织变革，重构业务流程

传统的企业是相对封闭的结构，与外界的接触多集中在供给环节和销售环节。在大数据时代，企业的围墙一一瓦解，各种资源向企业聚合以延伸企业的边界。企业在以平台为中心的生态体系中互惠共荣，过去相互独立的企业也以数据为纽带，开始共同编织一张大网，共生共长。

在内部价值链和外部环境的共同作用下，企业的组织模式开始发生剧烈的变化。传统的组织是高度专门化的，通过清晰的部门划分和明确的指挥链实现高效运转，企业内部的垂直边界和水平边界十分明确。而在大数据时代这一切都在改变，企业

① 参见德鲁克，《管理的实践》，机械工业出版社，2009 年版。

的垂直边界和水平边界开始消融，以任务或问题为导向的团队成为企业组织的常态，组织从原来的分工走向现在的合作，尊重员工、充分授权成为必然趋势；决策体系真正由原来的自上而下变成现在的自下而上，由原来的精英经验决策变成现在的数据驱动决策。

大数据正在掀起企业组织变革的新一轮狂潮，企业领导者要站在大数据的浪潮之巅，为组织变革做好准备。

第一节　大数据重塑企业内部价值链

提要：

1. 在大数据时代，每一个客户都能发出清晰的声音，单个客户的声音也越来越强，消费者的力量无可忽视。一个企业如果没有重视消费者的意识，必然会被竞争对手所替代，“以消费者为中心”成为必然的选择。
2. 依托海量的信息和大数据技术，新产品和服务产生的速度比历史上任何时期都要快，成长过程也爆发出前所未有的速度。
3. 当商品出现过度供给，企业价值链开始从生产驱动转向需求驱动。在大数据时代，客户实际上介入了企业，引导企业价值链转向趋于深度整合，驱动组织价值链智能化和柔性化。这在研发和设计、生产、供应、营销、售后服务等价值链环节都有所体现。

德鲁克问“谁是你的客户”，这个问题可以再深入地问下去：我的客户在什么地方，可以分为哪几个细分市场，各有什么需要……这些问题对一个负责任的管理者来说并不难回答。

再进一步，“你是在以客户为中心思考吗？你是在以客户为中心工作吗？”我们

早已进入供给过剩的时代，以客户为中心思考和工作的理念也传播了很久，但是我们一直在偏离。管理者习惯以市场份额、销售收入、销售增长率等指标作为衡量标准，但这些指标并非必然指向以客户为中心，甚至会出现违背客户意愿的行为。比如国内市场上，经常会出现挤压渠道、强制铺货、捆绑销售等短期行为，这些行为损害了企业的客户基础，不利于企业的长期发展。

脱离客户其实是企业成长过程中普遍的困扰。在创业阶段，运作团队小，企业的重心必然在客户，否则无法生存；当企业发展起来后，重心开始转移，逐步从面向客户转移到面向自己；当企业壮大之后，企业的重心离客户越来越远，很容易变得只关注自己，因为有太多的内部问题需要去解决。引导这几个阶段变化的主要原因是企业领导者的关注重心和工作重心从客户转移到了内部。

以客户为中心思考和工作，需要企业领导者改变自己的关注点，改变企业价值链的方向，从传统价值链向现代价值链转变。

传统价值链和现代价值链①

传统价值链以企业的资产和核心能力为中心，然后投入人、财、物，提供产品或服务，通过销售渠道，最终到达客户，如图 8-2 所示。传统的价值链在供给短缺的时代非常实用，直到目前仍然是许多大中型企业领导决策的基础。

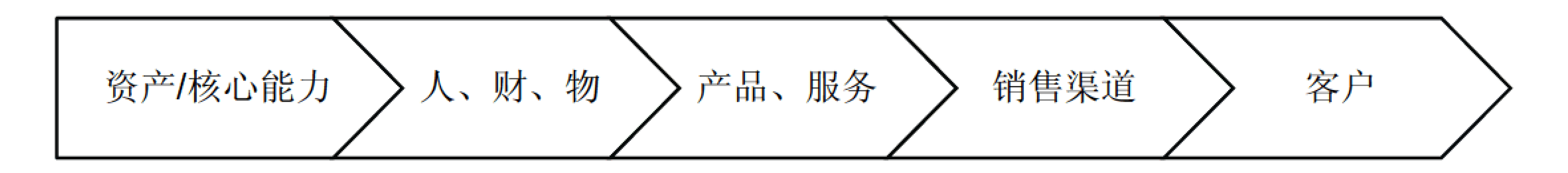


图 8-2　传统企业价值链

现代价值链以客户为中心，始于弄清楚客户的偏好，以何种方式满足客户的偏

① 参见亚德里安·斯莱沃斯基，《发现利润区》，中信出版社，2010 年版。

好，然后是最适合的产品或服务，最后才是人、财、物，以及支撑这些人、财、物的关键资产和核心能力，如图 8-3 所示。

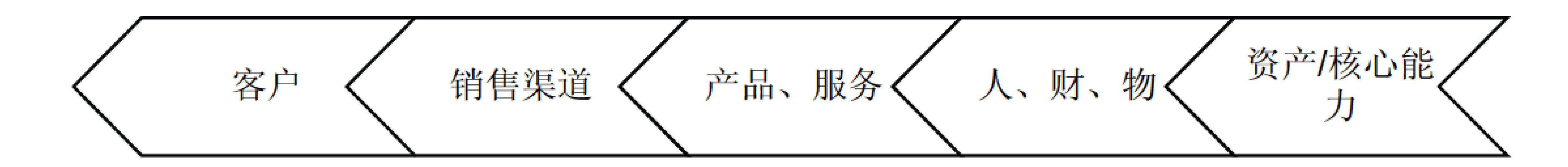


图 8-3 现代企业价值链

以客户为中心的现代价值链的概念几乎已成为普遍性的知识，企业领导者通过咨询顾问、培训班、专业书籍经常接触到，每次看都有所悟，似乎抓住了问题的关键，可是面对自己的企业时，却总觉得难以下手。企业内部没有形成以客户为中心的思考和工作的习惯，或者说缺乏这样的核心能力。

大数据带来业务和商业模式创新

在大数据时代，每个客户都能发出清晰的声音，每个客户的声音也越来越强，消费者的力量无可忽视。一个企业如果没有重视客户的意识，必然会被竞争对手所替代，"以客户为中心"成为必然的选择。企业领导者必须要思考如何有所作为，这将带来商业模式的根本变化。比如正在发生的典型商业模式：多边平台、定制化生产、依托海量信息开发新产品或服务……这一切无不取决于对客户的理解。

多边平台[①]将两个或更多的不同客户群体集合在一起，通过他们相互依赖的关系获取价值。平台必须能够同时吸引多个客户群体，并通过服务他们来创造价值。平台对某一客户群体的价值，取决于这个客户群体通过平台能够接触到什么样的目标客户，以及多少目标客户。大数据为平台的多边客户精确匹配提供必要的工具，促进平台多边的客户规模相互激发增长。淘宝就是这样的平台。2012 年 11 月 11 日，在这无中生有的"光棍节"，淘宝的商家和买家达成了 1.058 亿笔订单，销售额 191

① 参见亚历山大·奥斯特瓦德，《商业模式新生代》，机械工业出版社，2012 年版。

亿元，是 2011 年“光棍节”的 3.61 倍。谷歌的 AdWords 服务也是这样的平台，广告主可以在谷歌搜索页面上发布广告，当网民使用谷歌搜索引擎时，这些广告就会显示在搜索结果的旁边，谷歌的搜素业务收入也在持续强劲增长。在淘宝和谷歌的 AdWords 平台的背后，就是大数据的存储、处理、分析和应用能力。

定制化生产是面向需求，以订单驱动的生产方式，需要多个生产环节协同完成一个订单。以重庆美心集团与用友 U9 携手实现的“多组织协同下的大规模定制化生产”为例：美心集团以门业为主，产品标准化程度不高，产品品种繁多，生产工艺复杂多变，订货及变更非常频繁，对快速响应变化能力要求极高。用友 U9 系统帮助美心集团统一业务流程和整合信息，实现销售订单选配、网上订单和二次信用控制，集成条码系统、MES 系统、立体仓库系统和 PDM 系统，用柔性的系统和整合数据处理能力实现了定制化生产。

大多数企业都会开发新产品或服务，但依托海量的信息和大数据技术，新产品和服务的产生比历史上任何时期都要快，成长过程也爆发出前所未有的速度。微信对于腾讯而言就是这样的新产品。2010 年 11 月 20 日，微信正式立项，从立项到产品上线仅仅经过了 2 个月，2011 年 1 月 21 日，微信 1.0 for iPhone 发布，提供了免费短信功能，但市场反应平淡，因为运营商已经提供了丰富的短信套餐服务，正常用户每个月的包月短信根本消费不完。随后，微信 1.2 版迅速转向了图片分享。一般来说，人在有限的载体上没有耐心深度阅读，图片的吸引力会更大，于是微信 1.2 的主体功能变成了图片分享。但是数据反映用户对手机图片分享根本没有兴趣，无法构成一种基本的需求。随后微信 2.0 版从用户在手机上输入内容的便利性出发，将产品重心完全转入了语音通信工具，并把这作为一种重要指标，确立了微信 2.0 快速流行和传播的基调。微信 3.0 版运用大数据技术，提供了“查看附近的人”功能，成为微信的爆发点。从此微信再借用 QQ 邮箱和腾讯其他资源，向目标用户强力推广，用户数迅速突破了 2000 万大关，确立了对竞争对手的绝对优势。当用户超过 1 亿之后，微信结合大数据和熟人社交圈，于 4.0 版推出“朋友圈”，建立手机

上的熟人社交圈，开放 API 接口打造移动社交平台。微信 4.2 版再推视频通话功能。从此，微信确立了移动互联网时代基本生活方式的产品地位，大数据对微信的推广和爆发起到了决定性作用。

运用大数据，新的商业模式将以超乎人们想象的速度和形式出现，而这一切都被一条主线串起，就是企业价值链的彻底转变。

从理解客户开始重塑价值链

以客户为中心，从理解客户开始。大数据是理解客户的利器，在客户划分、客户定义、实时需求等方面，大数据都展现了它的精准和高效。

客户划分

运用大数据，客户划分可以从大众化、细分化变为微分化、个人化。

传统的市场一般分为大众市场、利基市场、细分市场等几个类别。在大众市场，企业的产品、服务、渠道、营销推广等聚焦于一个大范围的客户群，在这个客户群内，客户的需求基本相同，如个人计算机、运营商的通信服务；在利基市场，企业的产品、服务、渠道、营销推广针对某一特定市场的特定需求定制，如专供高档汽车的零部件、高档的跑车；在细分市场，企业的产品、服务、渠道、营销推广等在略有不同需求的市场群体之间会有所区别，如 SMH 不同品牌的手表、宝洁的三大品牌洗发水。

运用大数据，市场可以不再用上述的概念划分，直接实现微分化，甚至个人化，这得益于以下两个条件。

第一，丰富的数据量，为市场微分化、个人化提供了信息基础。传统的企业，除了电信、银行等个别行业，很少有具体到单个客户的详细数据。大数据时代，客户的数据在爆发性增长。“联网型组织”即开放内部的信息通道并通过互联网使客户和供应商参与进来的组织，优先具有了这个基础，如淘宝、京东商城、腾讯、百度、

新浪等互联网企业。传统企业中，奔驰、海尔、苏宁电器等也都在走向“联网型组织”。

第二，不断进步的大数据技术，让市场微分化、个人化并从中发现价值成为可能。亚马逊收集了上亿客户的单体客户行为：你搜索了什么，看了哪些产品的详细介绍，最终购买了什么产品，都会被亚马逊记录下来。而其他用户的历史购买行为在这里也将派上用场，成为相关推荐。因为大部分用户做购物决策的时候，很希望知道和自己有相似爱好的人看了什么，买了什么，有什么评价。

客户定义

运用大数据，客户定义从标准化向个性化转变。

传统的企业要了解客户，一般根据客户价值，或者通过问卷访问、小组访谈等市场调研技术去分析客户，通过一个或者几个维度来定义细分市场，给一个客户群一个标准化的面孔，再配套企业的产品、服务、渠道、营销推广。招商银行根据“一卡通”客户在银行的总资产量，将客户分为普通卡客户、金卡客户和金葵花卡 VIP 客户；中国移动根据客户价值、客户行为和需求，将客户分为追求成就的高价值全球通客户、年轻有潜力的动感地带客户和享受实惠相对低端的神州行客户。

大数据给出的客户定义，不再是一个群体标准化的面孔，而是立体、全面的客户形象。这个形象由两方面数据组成：一是结构化的交易数据，如消费水平、消费频次、生命周期等；二是非结构化的交互数据，如文本、图片、多媒体等，这些数据以远大于交易数据的增长速度呈指数增长。利用大数据技术，对交易数据和交互数据综合分析，客户的定义就成为一个个丰满、立体的个性化形象，而不再是抽象出来的标准面孔，这无疑更精准地反映了客户的需求。

部分零售企业走到了利用 “大数据”的前沿。它们运用“情感分析”技巧，发掘使用社交媒介的消费者产生的海量数据流，及时掌握客户的情感变动，适时调整推荐产品及对应的活动，提升了商品的周转速度和毛利空间。

谷歌是利用大数据定义客户的先驱，通过免费的软件及服务来更精确地理解客

户行为和习惯。谷歌提供的免费软件越多，对客户的理解就越深刻，如谷歌图片、谷歌音乐、谷歌邮箱、谷歌视频等都为企业提供了从不同角度理解客户的机会。谷歌在精确理解客户的基础上再向企业提供精准的广告服务，创造了高利润的商业模式。

实时需求

客户的实时行为倾向是最有效的客户需求信息，但这样的信息稍纵即逝，用传统的方法很难捕捉。大数据能准确获取客户实时的个性化信息，帮助企业做出高效、有针对性的决策。

传统的零售企业稍加改进就可以获取客户的实时行为信息。比如，在购物车上装上传感器，可以跟踪客户的行进路线，获取在不同位置的停留时间，以及实际购买的物品。利用这些信息，有助于卖场及时调整货架展陈、上架商品，提高销售额和利润率。

互联网零售企业不仅可以获取实时信息，还可以实时分析用户行为、实时调整公司经营策略。通过互联网点击流可以实时跟踪客户的行为、更新他们的偏好，并建立客户行为的可能性模型，实时推荐优选商品，提供省钱的奖励性计划，使整个销售流程圆满结束。

线下商家可以通过实时的位置信息向周边客户推送最新的优惠活动。智能手机在迅速普及，基于手机位置信息的应用将会蓬勃发展。当智能手机用户靠近一个运动服装店时，这个服装店可以向周边的智能手机用户推送最新的限时打折消息以提高销量。基于位置信息的社交客户端也可能被用来开展此类的推广，微信上已有商家开始这项工作了。

大数据驱动价值链向智能化和柔性化方向发展

当商品出现过度供给后，企业价值链开始从生产驱动转向需求驱动。在大数据时代，客户实际上介入了企业，引导企业价值链趋于深度整合，驱动组织价值链智

能化和柔性化。这在研发和设计、生产、供应、营销、售后服务等价值链环节都有所体现。

营销和售后服务的变化

运用大数据的企业可以定位微分化、个人化的客户，实时、全面地把握客户的需求特征，以传统企业前所未有的智能和柔性程度高效营销。Kindle 的推送功能就具有这样的特征。用户把想要在 Kindle 上阅读的文档以附件的形式发送到亚马逊分配给其的一个邮箱地址，那个邮箱地址不是真的可以用来收发邮件，而是每台 Kindle 专属的云存储空间。当用户的 Kindle 链接上网络的时候，就会自动下载用户没有下载过的文档，用户也可以选择在任意时间反复下载其在云存储空间中的任意文档。

在售后服务领域，大数据同样展现了智能化的潜力。企业通过远程智能监测、远程辅导等方式，可以有效降低人员投入，提高服务质量和服务效率。比如，设备制造商如电梯制造商、飞机制造商、机床制造商等，在售出的设备中植入传感器，传感器实时记录设备的运行情况，并回传给设备制造商，设备制造商就能立刻获知设备运行中的问题，迅速做出诊断，无需到现场即可远程指导企业维护调整。在远程智能维护中，长期积累的数据是设备保证系统有效运转的核心。

生产的变化

大数据对生产流程的改变包含以下方面：柔性化生产，满足个性化需求；运用模拟技术，降低生产风险；远程实时监控，改善操作环境。

未来当每个消费者的声音越来越强，消费者群体的力量越来越大的时候，价值链第一推动力就必定来自于消费者，那时“定制”商业模式会是主流。个性化需求要求的是多品种、小批量和快速反应，率先实现柔性生产的企业将具有极大的竞争优势。柔性化生产的重大难题是成本问题，技术的进步，尤其是信息技术在生产中的大量应用，使柔性化生产的成本大幅下降，基于个性化需求的生产计划实现柔性

化，同时生产线流程也被信息技术加以改进以实现柔性化。柔性化的基础是数据的获取、传输、运用，而这在汽车、家具等制造行业已有深入的应用。

生产过程中难免遇到种种风险，可能会导致巨额的损失。基于大量历史经验数据和规律，可以对生产过程进行模拟分析，提前发现风险，做好防范措施。

当生产环境不适于人工操作，或人工成本大幅上升时，机器将代替人力。在这些机器中植入传感器，依赖传感器回传的数据，可以远程监控设备运营，操作人员远程操作指导机器工作。

供应链的变化

供应链管理最关键的环节就是市场需求预测。大数据可以高效分析大量个性化的需求，结合各种辅助信息，通过合理的预测，灵活、适时地安排供应链各个环节的工作。

一般供应商可以通过自身积累的数据，改进需求预测，安排供应计划。当供应链上下游的数据变得透明时，供应商将能够更合理地安排生产和供应。比如，当生产消费品的企业获取到零售商的数据、生产零部件的企业获取到设备生产商的数据后，可以更合理地安排物流、生产、原材料等供应链环节。除此之外，还可以创新整合上下游以外的数据来创造价值。比如，一家全球性饮料企业将外部合作伙伴的每日天气信息集成，进入其需求和存货规划流程，通过分析特定日子的温度、降水和日照时间三个数据点，该企业减少了在欧洲一个关键市场的存货量，同时使预测准确度提高了大约 5%[①]。

研发和设计的变化

在过去，研发设计、供应、生产、销售、服务等环节的信息都局限在相互独立的部门中。因为对结构和非结构化数据处理技术的进步，上述各个环节的信息被有

---

① 参见《麦肯锡季刊》中文版(china.mckinseyquarterly.com)。

效地整合起来，研发设计人员可以方便地从其他环节的数据中及时提取到有价值的信息，如新产品在生产过程遇到的问题、新产品的销售状况、客户对新产品的反映等。这些信息原来需要通过定期的分析才能获取，而运用大数据，研发设计人员可以实时得到这些数据，从中提取出有价值的信息，及时改进产品的设计方案，加速产品更新的进程和对客户的响应。

海尔构建了内部市场链的研发及营销协同机制，研发部、企划部、市场部和售后部多边参与，基于多维度的数据构建了新产品市场检测体系，形成一个迅速了解市场信息的开放系统，为企业不断优化产品、推陈出新提供了保障。

UC 浏览器牢牢占据了亚洲市场四成的份额。UC 浏览器能够取得这样的成绩，是因为其一直秉持着立足市场为客户创造价值的研发理念，智能适应屏幕排版、夜间模式、语音等优秀的功能，都是基于对客户行为数据和反馈信息的长期跟踪而开发的。UC 优视的 CEO 俞永福说："在全球化过程中，移动互联网企业必须做到'全球化思考，本地化执行'，在一些重点区域市场的开拓中，不能仅仅做在产品上的语言翻译、横向移植，必须对当地的文化有充分的了解。"这实际上就是对当地的客户行为特征数据的充分掌握和分析。

如果企业的买卖关系都已经电子商务化了，大数据对企业价值链的影响将更全面地显现出来。阿里巴巴集团总参谋长曾鸣说："当互联网继续推动到整个价值链的各个环节，信息都能在网络、不同的 Player 之间实时协同分享的时候，那个时候电子商务才真正发挥出它的威力，它是一种全链条的价值再造过程，是一个价值创新的过程"。

第二节　大数据改变组织的外部边界

提要：

1．大数据给企业提供了便利的工具，创造整合外部资源的机会，降

低整合外部资源的成本。接近客户的企业因为能够更精准地接触客户、理解客户，有很强的话语权，得以聚合周边的资源以延伸企业的边界。

2.“平台+商家”的双层结构生态体系中，所有权与使用权出现了分离，大量的商业流程被平台上流动的数据驱动，并在企业之间、企业与消费者之间灵活组合。以平台为中心、以企业和消费者为两翼的平台生态体系成为新的组织形态。

3. 数据不同于有形资产：有形资产分享的越多，自己拥有的越少；数据分享的越多，产生的越多，使用的人越多，其价值越大。因此，数据具有天然的公用性和价值性。当数据成为一项重要的资产后，原来基于资产专用性的企业边界也变得模糊。

科斯用企业内部的管理成本和外部的交易成本来解释企业的边界。当内部的管理成本小于外部的交易成本时，企业倾向于通过内部管理配置资源；当内部的管理成本大于外部的交易成本时，企业倾向于通过外部市场配置资源。企业扩张会带来组织内部的管理成本增加，当内部的管理成本等于外部的交易成本时，企业将停止扩张[①]。

大数据让企业的内部管理成本和外部交易成本都大幅下降，这两者下降的速度在不同的企业是不平衡的，有些企业管理成本下降的速度快于交易成本，有些企业交易成本下降的速度快于管理成本。因此，企业的组织形式将朝着两个方向发展：前者的企业规模将不断扩大，企业规模的记录将不断被打破，典型发展路径是掌握客户的企业沿着产业链向上整合；后者的企业规模将变得更小，更依赖于外部的资源，典型发展路径就是依托一个平台实现资源的快速、低成本交换。

① 参见黄家明，《交易费用理论：从科斯到威廉姆森》，合肥工业大学学报（社会科学版），2000 年。

大数据推动资源聚合，延伸企业的边界

整合外部资源是企业重要的战略选择，一般需要比较严格的条件，要么在市场上或者在产业链上有很强的话语权，要么付出巨大的经济代价。大数据给企业提供了便利的工具，创造整合外部资源的机会，降低整合外部资源的成本。接近客户的企业因为能够更精准地接触客户、理解客户，有很强的话语权，得以聚合周边的资源以延伸企业的边界。

2000 年，宝洁新任 CEO 雷福礼临危受命，为重振宝洁将创新作为公司的核心。其关键因素之一就是连接和发展战略，旨在通过外部伙伴关系促进内部的研发工作，计划将公司与外部伙伴的创新工作提高到总研发量的 50%。为了连接企业内部资源和外部伙伴的研发活动，宝洁建立了三个桥梁：技术创业家，他们是来自宝洁内部的高级科学家，与外部的大学或其他研究人员建立良好的关系，还扮演“猎人”角色，寻找外部的解决方案以解决内部的挑战；互联网平台，宝洁把一些自己研究上的难题，通过该平台暴露给全球各地宝洁以外的科学家，成功开发出解决方案就可以获得奖励；退休专家，宝洁通过 YourEncore.com 网站从退休专家那里征求知识。在过去，这些桥梁很难建立，而有了互联网和大数据，宝洁能够很方便地获得大量的外部资源，方便地管理、运用这些外部资源。2007 年宝洁就完成了既定目标，研发生产率也大幅提升了 85%。

客户参与产品研发和推广也是互联网公司的通用做法。互联网公司开发一款新产品，一般都会邀请客户测试，或者把产品放到玩家客户群中，让他们使用、提意见，把客户的设计体验融入到产品研发中。这些客户很乐意无偿提供自己的才智，在帮助企业改进产品的过程中获得乐趣，当这些产品符合自己的想法或者满足了自己的需求时，又很乐意担任这些产品的免费推广员，微信、新浪微博、UC 浏览器等等都是这样发展起来的。

接近客户的企业利用客户赋予的话语权，可通过以下几种方式聚合资源以延伸

企业的边界。

1. 将供应商或经销商整合进自己的价值链。当能够利用数据交互提高整合后的效率，或者能够扩大企业收益基础时，这些接近客户的企业就有了整合的动力。鸿海集团最厉害的就是有强大的垂直整合的产业结构：接近重要策略客户，与客户实时联动研发、设计、测试、发布新产品；建立全球物流追踪系统，依据客户需求及时调整存货，客户要货时有货，不要货时零库存；当市场低落，产品利润比较低时，上游零部件商就可以补贴下游组装厂，制造的成本永远比竞争对手低，毛利率永远比竞争对手高。

2. 通过供应链优化整合上游企业。这种整合情况多发生在供应商、经销商不具有独特的核心价值，但通过数据交换可以提高效率的时候。苹果的核心力在其创新能力和品牌影响力，制造相对来说就没有核心价值，因此苹果只是选择、协调、监控供应商，借助完善的零部件数据、生产过程数据、质量监测数据等优化供应链的效率，以降低生产成本和供应风险。

3. 聚集其他资源创造出新的商业模式。中国移动的彩铃业务就是通信行业聚集音乐行业资源产生的新商业模式，它之所以能够成功，除了客户需求外，关键在于中国移动有具体到每个客户的信息、订购关系，有基于每个客户订购关系的支付手段。谷歌正试图利用大数据创造一个个绚丽的新商业模式。如谷歌无人驾驶汽车通过摄像机、雷达传感器和激光测距仪来“看到”其他车辆，并使用详细的地图来进行导航。这基于谷歌强大的数据中心来实现的：数据中心收集了大量的手动驾驶车辆的信息，并对这些信息进行了处理转换。谷歌还在研究模拟人脑：谷歌的科学家将 1.6 万片计算机， 处理器连接起来，创造了全球最大的神经网络之一，让它们在互联网中“自学成才”。这个神经网络已经依靠自学认出了猫咪。这项研究是大数据远大前景的代表，充分利用了下滑的计算资源成本，以及日益增多的庞大数据中心。此外该技术还大力推动了众多领域的进步，包括机器视觉与感知、语音识别、语言翻译等，将创造出一个又一个突破人们想象力的商业模式。

以平台为中心的生态体系共生共长

以平台为中心的生态体系是大数据时代的主流商业模式，这些平台有优秀的消费者资源和商家资源的吸纳能力和服务能力，平台和平台上的商家互惠互利，共生共长。阿里巴巴是这样的平台，腾讯是这样的平台，新浪微博也有成为平台的潜力，传统的家电连锁企业苏宁电器也在努力成为平台。

平台生态体系具有典型的网络效应，一旦突破临界点，平台的生态体系将不断向外发育，飞速成长。如果网络中只有少数用户，他们不仅要承担高昂的运营成本，而且只能与数量有限的用户交互，价值很低。假设淘宝只有很少的买方、卖方，买方就买不到需要的商品，卖方也挣不到足够的利润，淘宝平台也就无法持续运营。随着用户数量的增加，这种不利于规模经济的情况将不断得到改善，所有买方和卖方都可能从扩大的网络规模中获得更大的价值。卖方数量增加就能吸引更多的买方，因为他们能很方便地买到需要的商品，而且可以“货比三家”，获得较好的性价比；买方数量增加就会有更多的卖方进驻，因为卖方的利基在扩大，可以获得更多的收益。一旦卖方或买方的数量突破临界点，对买方或者卖方来说边际效应得到显著提高，更多的买方和卖方会依附到这个平台的生态体系，整个生态体系也就不断向外扩张，蚕食那些孤立的领地，进入“赢家通吃”的阶段。

淘宝在建立的初期采用免费的策略，迅速突破了生态发展的临界点，不到两年用户数就超过了 700 万。2003 年 8 月 17 日，淘宝网对外宣布，前 10 万名经过身份认证并在淘宝上有过一次买卖经历的会员，将享受 3 年内不收取交易服务费的优惠，使得淘宝在与收费的 eBay 易趣竞争中极具吸引力，吸引了大量卖家迁移，很快 eBay 易趣在中国就没落了，而现在的其他电子商务网站对于淘宝也只是难以望其项背。

数据是平台生态体系的核心资产。阿里巴巴将自己定位为“数据分享的第一平台”，让大数据流动到有需要的个人或企业那里，挖掘出它的价值。“聚石塔”就是

阿里巴巴的一款大数据产品。聚石塔向商家及第三方服务商提供阿里集团的云资源，如云主机、云存储等，使商家和服务商的数据与业务流程能够实现云化，拥有足够的 IT 可靠性与灵活性。聚石塔具有数据推送的功能，以前商家与淘宝网通过公网数据接口的方式实现数据交互，效率比较低，且容易出错；在聚石塔的平台上，基于阿里巴巴内网推送的数据能够快速、高效地推送到商家的数据库里。聚石塔还具有数据集成的功能，规划中的聚石塔会形成统一的电商数据总线和接口标准，各个软件系统都与这一数据总线协同，同时通过订单状态的数据标准，使订单在各个软件之间快速流转。最终，聚石塔将通过处理、整合、开放和共享会员信息、商品信息以及交易信息等数据，实现商家所使用的 IT 服务商的各类 IT 系统间的彼此连通，获得高效、全面的大数据服务。

建立强大的平台生态体系是众多企业的梦想，但并非所有企业都能做平台，现实中大量存在的是围绕平台的小而美的企业，它们与平台共存共荣。马云在 2012 年网商大会闭幕演讲时讲到电子商务的一个重要趋势就是“小就是美，Small is beautiful”，“……阿里将全面推出双百万战略，全力培养一百万家年营业额过一百万的网店，……这种规模是最有味道，最好的，只要你持久长，小企业因为你幸福，因为你好这口，你就会有不断的创新。”

“小就是美”成为趋势主要受两个因素驱动：第一，平台生态体系为企业提供了丰富的个性化需求信息，匹配个性化需求的企业将会优先得到发展，获得较高的利润率。满足个性化需求需要柔性化定制，实现个性化制造，但柔性化定制也限制了规模的扩大，在平台上很难出现一个一统江湖的企业。第二，平台生态体系上的企业依赖平台获取客户、开展营销、管理运营，这些工作原来需要企业投入很多的人力、财力和物力等资源，而依托平台的大数据能力，这些工作都将被简化，只需精心做好自己的产品，就能依托平台实现快速增长，相较于以前的同等收入级别的同类企业，人员规模、资产规模的需求要小得多。

2012 年十佳网商中，美国 Ever-Pretty Garment(艾娃贝蒂)企业创始人 Anna

女士，主营婚纱礼服产品，从 eBay 网店起步，后来通过阿里巴巴平台走向了全球市场。虽然企业规模小、资源不足，但是通过电子商务，她为客户创造了独特且舒适的购物体验。徐长应是一名小产品营销专家，他创立的北京银曼是一家集家居清洁，美容护肤，健康养生的研发、生产、销售为一体的综合企业，2008 年开始做电子商务渠道，目前旗下有宣琪、凡茜、足季、浴见知己、软么么 5 个品牌，其产品虽小，但却开创了从分销到零售，从小而美的品牌制造到品牌孵化的互联网模式，在网络上多次创下单品销售上亿元的纪录。

“平台+商家”的双层结构在信息产业中的应用将更加普遍。早在 2000 年底，中国移动就推出了“移动梦网”创业计划，改变过去由电信运营商独自服务用户的历史，用“移动梦网”这个平台，集合第三方服务、内容供应商，主要向手机用户提供基于短、彩信的移动数据业务。如今的移动互联网时代，大量的第三方开发商为苹果的 App Store 开发应用，截至 2012 年 9 月，App Store 的应用数已经超过 70 万。中国移动也开发了移动 MM 应用商店，目前也有了 15 万左右的应用数。

“平台+商家”的双层结构，对客户来说有了前所未有的丰富选择，对企业来说降低了 IT 设施的投资成本，以及各种运营管理成本，其必然会成为未来的主流商业模式。

在过去，资产的专用性为组织确定了明确的边界。“平台+商家”的双层结构生态体系中，所有权与使用权出现了分离，大量的商业流程被平台上流动的数据驱动，并在企业之间、企业与消费者之间灵活组合。以平台为中心的、以企业和消费者为两翼的平台生态体系成为新的组织形态。

大数据将相互独立的企业串成一张大网

人与人之间的网络化联系在不断地被深化。基于交通网的物流是最底层的网络化，马道和马车、公路和汽车、铁路和火车、机场和飞机一次又一次地帮助人们突破与外界交往的空间，人类生产和生活的物资被快速地运送到更远的距离；基于金

融网的资金流是第二层的网络化，钱庄和飞钱、银行和纸币、股市和股票、债市和债券……，每一种新的资金形式出现都加快了资金流转的速度，扩大了资金到达的范围，资金流最初基于物流产生，后来又脱离物流形成纯粹的资金流，如金融衍生品的买卖；基于电话、传真、互联网的信息流是第三层的网络化，信息流提升了沟通的效率，促进了物流和资金流的流通，信息与信息之间的交互会产生更多的信息，QQ、Facebook 上的内容多是基于虚拟的信息交流产生的。

大数据丰富了信息流即第三层网络化的内容，大数据技术提升了信息网络化的价值，相互独立的企业被紧密地联系在一起，所有的企业都成为网络的一个节点，共同为这个网络贡献数据，也从这个网络上获取数据。

数据不同于有形资产。有形资产分享的越多，自己拥有的越少；而数据分享的越多，产生的越多，使用的人越多，其价值也就越大。因此，数据具有天然的公用性和价值性。当数据成为一项重要的资产后，原来基于资产专用性的企业边界也变得模糊了。

将一些私有数据公用化将产生巨大的价值。很多医院目前都已经实现了信息化管理，拥有客户的年龄、病情、用药、手术、费用、治疗效果等数据，这些数据成为医院的私有资产。如果这些信息被集合在一起公用化，医药企业、医生、病人、政府围绕这些数据形成相互支持的网络，将有极高的医疗价值，大幅度降低社会的成本。当很多医院的客户信息被集合在一起后，关于病情、用药、手术、费用、治疗效果之间的关系就有了以下几方面的价值：医药企业利用这些信息可以降低研发成本，提高研发效率；医生利用这些信息可以降低医疗事故发生率，提高医治水平；政府利用这些信息可以很容易评估过度治疗的现象，惩罚过度用药的医院，降低病人医疗费用和国家医保费用负担，合理规划医保投入。[①]这些数据有极大的社会福利效应，政府应该积极推动公用化，同时处理好医院的利益和知识产权间的关系。

---

① 参见麦肯锡，《Big data: The next frontier for innovation, competition, and productivity》，2011 年。

企业应该重新评估自己的数据管理政策，让一部分数据公开化，做到在为网络创造价值的同时，自己也获取更多的收入。腾讯在 2011 年前是一个相对封闭的企业，不开放自己的平台，利用用户规模优势，模仿其他企业的创新，迅速超越、挤垮对手，激起了互联网上声势浩大的反对浪潮。2011 年认识到开放可以放大自身价值后，腾讯将开放平台列为长期战略，目前已经开放了 Qzone、财付通、微博客，推出了类似 Facebook like 功能的 QQ 空间的“喜欢”，并且为了开放战略全资收购了著名的开源软件 discuz 的所属企业康盛创想。腾讯开放的领域对其自身而言，是“增量市场”，既不会影响腾讯自身的现有业务，又能让腾讯分享其他企业依靠其平台获得的增长和收益。

拥有数据的企业如果主动开放，审慎地与其他企业建立数据和收益的分享机制，将会创造出巨大的商业空间，网络上的数据流量及其价值也将会继续呈指数级增长。互联网企业已经率先开放，运营商、保险、银行、医院等等传统的拥有大量客户数据的企业是否也能走出各自独特的数据开放之路呢？

任何网络都有突破的临界点，大数据网络的临界点一旦突破，人的衣、食、住、行、医、娱等都可以从这个网络上获得满足，所有的企业都在这张大网上共生共长。大数据是穿透企业围墙的利剑，无论你是否愿意，你都将在这个网络上生存。与其等待，不如主动拥抱！

第三节　大数据推动企业组织管理变革

提要：

1. 在大数据时代，也许可以提供解决大企业病的药方。不同专业、不同类型的数据都能被广泛获取，在组织内有序传播，被合理地解读，组织内部的透明度和沟通效率大幅提升；同时管理层级进

一步扁平化；这些特征带来组织管理模式的剧烈变化。

2. 有了大数据资源和大数据技术，一个小团队发挥的作用将超乎想象。腾讯、谷歌等互联网企业都采用了小团队的管理方式，以产品经理为核心的产品团队，贡献了很多创新的、具有巨大市场前景的业务。

传统企业组织发展一般会经历以下的过程：诞生时是几个人的小团队，分工不明确，每个人都承担着多种任务；随着业务成长、规模扩大，开始专业化分工，建立职能制，相同的任务组成专业部门以提高工作效率，形成层级，规定谁向谁汇报工作，保证组织的政策、策略上通下达；当有多种业务或者在多个地区经营时，原有的职能制决策负担过重，不能有效行使决策权，就出现了事业部制；再往前发展就会出现矩阵制……在组织进化的过程中，企业的员工会越来越多，沟通效率越来越低，“以邻为壑”的现象时有发生，同时各层级的管理人员也大幅增加。现在，几乎每个大企业的领导都在与大企业病斗争。

在大数据时代，也许可以提供解决此类痼疾的药方。不同专业、不同类型的数据都能被广泛获取，在组织内有序传播，被合理地解读，组织内部的透明度和沟通效率大幅提升；同时一些可以用运算解决的决策工作也被数据替代，不再需要那么多的管理人员。这些特征带来了组织管理模式的剧烈变化。

大数据推动分工走向合工

大数据时代，数据传递透明、迅速、全面，企业近似于一个全息的有机体。有机体有血液，依托血管网络迅速完整地传递到身体的每个角落。而企业的每个岗位都会产生数据，可经网络实时、全面地流转。大数据相对于企业就像血液相对于有机体，信息网络就像血管；有机体的一个动作需要全身肌肉实时配合，企业的一项工作，也能基于信息网络和数据迅速调动企业内的各项资源。

在类似有机体的企业内，每个人的工作都是为了整体的成功。传统的分工是为了分解某项工作，提高工作效率，在某个部门内、在某个岗位很难看到整体的目标。运用大数据，企业的垂直和横向都能够实现实时、高效的沟通，每个人都能够通过正规渠道及时了解企业各方面的动态信息，在相互透明的环境下，每个人都关注整体的成功，互相协作。就像人一样，各个部分的动作都是为了完成大脑的某个指令。部门的界限和层级分工会仍然存在，但是跨越界限的合作必然将成为组织运作的主流。

大数据重组组织的垂直边界

大数据实现了信息的扁平化。互联网普及以来，多数企业都或多或少利用了信息化，如 OA 系统、CRM 系统、ERP 系统等等，这些系统提高了企业的工作效率和沟通效率，但是仍有几个问题需要解决：

1. 多种数据分散在不同的系统，如何去整合？
2. 如何实时获得数据、把握经营状况？
3. 如何自动分析系统中的数据，提高分析的效率？

这些问题都需要依赖大数据的技术去解决。支持结构化和非结构化数据的数据中心，将是大企业的核心资产；小企业也可以利用公共的数据中心。数据中心支持实时获得经营数据，并通过多种终端推送到需要这些信息的岗位；大数据的算法也被用来自动分析数据，在海量数据基础上形成分析报告。

信息的扁平化支持业务流程的扁平化。运用大数据，中间环节将被压缩，业务流程将有巨大的效率提升空间。比如，传统的电信运营商在制定一线营销任务、评估完成情况时，一般要经过以下几个流程：

1. 企业每月召开生产经营分析会，确定下阶段工作任务重点；
2. 根据企业工作任务重点，各部门将本部门工作分解成一个个项目；
3. 针对一个个项目，提取目标客户，下发给一线营销人员；

4. 一线营销人员根据项目任务，针对目标客户推广；

5. 管理人员按天或者按周跟踪完成情况。

当数据中心建立后，每个客户都有订购、终端、行为、服务等方面的信息。分析这些信息，运营商就可以事先预测每个客户还需要哪些业务、哪些服务，甚至可以实时更新并利用客户的信息。比如，当用户产生大量流量而没有办理流量套餐时，运营商就可以通过短信提醒，推荐优惠的流量套餐，在给客户带来优惠的同时也给运营商提供了稳定的收入。运用大数据中心，电信运营商在制定一线营销任务时，可以将流程做到非常简化：

1. 企业事先研究数据中心对客户需求分析的准确性，以及对一线营销支撑情况；

2. 各部门事先围绕数据中心，优化分析模型，对每个客户形成业务和服务“拼盘”；

3. 一线营销人员根据“拼盘”推广业务，管理人员可实时获得完成情况。

由上可见，新的流程和原流程相比，有以下三个优点：第一，新流程的沟通环节比原流程少，信息衰减少；第二，新流程的支撑工作可以预先进行处理，营销结果也可实时获得，整个流程的时间跨度短；第三，新的业务流程以客户为中心，改变了原流程的以管理为中心的方式。

业务流程的扁平化需要组织的扁平化。传统的业务流程基于职能层级制定，带有许多职能层级的色彩，这些对于扁平化的业务流程显得多余，也必然带来干扰。当扁平化业务流程的优点呈现在眼前时，组织领导应该勇于改革，删减多余的部分，设计扁平化的组织结构，从管理控制为主向支撑服务为主转变。

大数据融合组织的水平边界

在传统企业内，各部门专业不一样，信息格式也不一样：市场的信息多是结构化数据和多媒体数据，客服的信息多是文本数据，研发的信息则多种多样……各个模

块相互独立，互不了解，各部门之间的“以邻为壑”的现象时有发生。这样做造成了企业对市场反应迟钝，内部协同和沟通成本过高，以及部门工作重点和组织工作重点发生偏离。

企业家和管理学家已经为跨部门的问题的解决设计了多种方案，形成了矩阵型组织、跨职能团队等等。这些组织形式在增进相互了解的同时，又产生互相推诿、职责不清等问题，实质上并没有有效解决跨部门的问题。

大数据时代，部门间信息变得透明，信息能实时传递。因为数据存储技术的进步，市场、客服、生产、供应、研发等不同部门的信息第一次被有效整合起来，数据之间还建立了实时的联系，共同组成完整、全面的信息流。一个客户看到促销活动购买了一个产品，又向客服部门反馈了不满或者要求提供后续服务，这都在系统中被实时记录下来。向上溯源，可以看到这个产品的生产批次、质检人员，研发人员又可以把这些信息和研发时的参数进行比对……这些信息相互之间的联系可以被实时获取并分析，部门间信息的透明在技术上已经没有障碍了。

透明实时的信息促进部门间边界的融合。以前部门间信息传递是单向的，线性发生的，时间跨度长，信息被按照本部门的利益进行筛选。为了解决这个问题，有的企业建立了服务于整个组织而不是服务于某个部门的分析团队，这有利于解决部门间的矛盾，但是也人为造成了分析团队与业务脱离、分析结果有可能浮于表面的问题。当数据是全方位、实时产生并相互关联的时候，各部门的协作就像互联网一样并发、实时地协同；透明的信息促使信息节点之间互相监督，“以邻为壑”的节点将会被孤立，部门将自动服务于整个组织。同时因为实时并发，部门之间协同效率也将今非昔比。

大数据也会推动部门重组。部门间边界融合的下一步就是部门重组，以客户为中心组织价值链，需要灵活、实时、动态协同。按专业分工的部门总会存在影响效率的问题，具有未来眼光的企业将会重组部门、改变流程，最有可能形成网状的组织结构，让每个节点都能获得足够的信息支撑，并承担决策的功能，来适应流动的、

非结构化的数据。

小团队成为组织的常态

有了大数据资源和大数据技术，一个小团队发挥的作用将超出人们的原有理解。腾讯、谷歌等互联网企业都采用了小团队的管理方式，以产品经理为核心的产品团队，贡献了很多创新的、具有巨大市场前景的业务。

腾讯的微信就产生于这样的优秀团队，在短短一年半的时间内用户数就超过了一亿。当然优秀的产品基因是微信快速传播的基础，但是没有大数据的支持，微信的发展不会如此迅速。微信在发展过程中，充分利用了 QQ 客户端和 QQ 邮箱等产品的既有客户群：微信可以直接使用 QQ 号登录，可以添加 QQ 好友；登录 QQ 邮箱，就有可能看到微信的广告，QQ 好友的微信。庞大的腾讯客户基础以及客户关系的数据，是微信团队能够成功的必要条件。

谷歌采用的也是一种小团队管理方式，这种小团队方式有利于高效的创新、高效率的工作，相当于在大企业内有了创业企业的良好氛围。谷歌的前 CEO 施密特说："小团队管理方式主要有三个好处：一是它能够增加尝试的可能性，让我们不断尝试尽量多的新生事物，这样我们成功的几率就比较大；二是能够给我们的员工更多的主人翁责任感，让他们觉得不是在一家大企业工作，在开发过程中让他们觉得自己拥有决定方向的自主权，同时又可以为用户来服务；三是能够降低团队内部协调的成本。"谷歌几乎每个项目都是小组项目，每个小组之间都必须进行交流合作，小规模的团队让交流简单、有效。谷歌是"大数据"技术的奠基人，一个个小团队之所以能够发挥巨大的作用，背后离不开谷歌丰富的客户数据以及大数据分析技术的支持。

大数据使企业有序地充分授权

授权是很多企业管理者纠结的问题，一收就死，一放就散；不放企业就无法应

对多变的市场，不收企业就变成一盘散沙，很多企业管理者就在这二者之间来回摇摆。大数据时代，客户已经完全介入企业的行为，授权给接触客户的一线是必然趋势。大数据也是实现有序授权的有效工具，运用大数据，因授权导致失控的现象将一去不复返。决策体系真正由原来的自上而下变成现在的自下而上，由原来的精英经验决策变成现在的数据驱动决策。

海尔“人单合一双赢”模式走过了油水分离到水乳交融的过程，通过充分授权再造了海尔的组织结构。“人”是员工，“单”不仅是订单，还是市场目标，也就是一种广义的用户。用户深刻影响企业，每个一线员工不仅是一个员工，还是一个信息终端，他们了解客户需求，创造客户需求，将需求传输到信息中心，企业再根据需求提供足够的资源支持。以前员工听企业的，现在变成了员工听用户的，企业听员工的，其中数据能够透明、迅速、全面传递是保证模式顺畅的关键和前提[①]。

海尔的组织形式从正三角（见图 8-4）变成了倒三角（见图 8-5）。销售组织结构原来是全国一省一市一县的模式，现在是扁平化的形式，全部关注县一级。全国 2800 个县，海尔有 1000 多个一级经营体，分别对应这些县。员工在最上面发现需求和创造需求，领导在最下面由原来的指挥者变成支持者一同为客户的需求服务。

原来一个人只能管 8 个人，现在一线拥有 2000 多个经营体，每个经营体面向客户需求自组织、自驱动、自创新、自运转，极大提高了一线员工的活力和创新力，更好地适应了互联网时代营销碎片化和需求个性化的特点。

在这样的企业，企业的一切节奏都自发加快，不需要等待自上而下的推动。在客户需求的牵引下，企业中的每个人都不断地获取客户需求信息，改善产品和服务，否则将被不断变化的市场所淘汰。大数据让企业有序地充分授权，充分授权让企业自发地适应客户成长。企业家应该做好引导，让企业面向未来，走上自我进化、自我超越、自我发展的健康道路。

---

① 参见张瑞敏，《管理创新与 IT》，SAP 中国商业同略会暨 SAP 全球技术研发者大会演讲，2011 年。

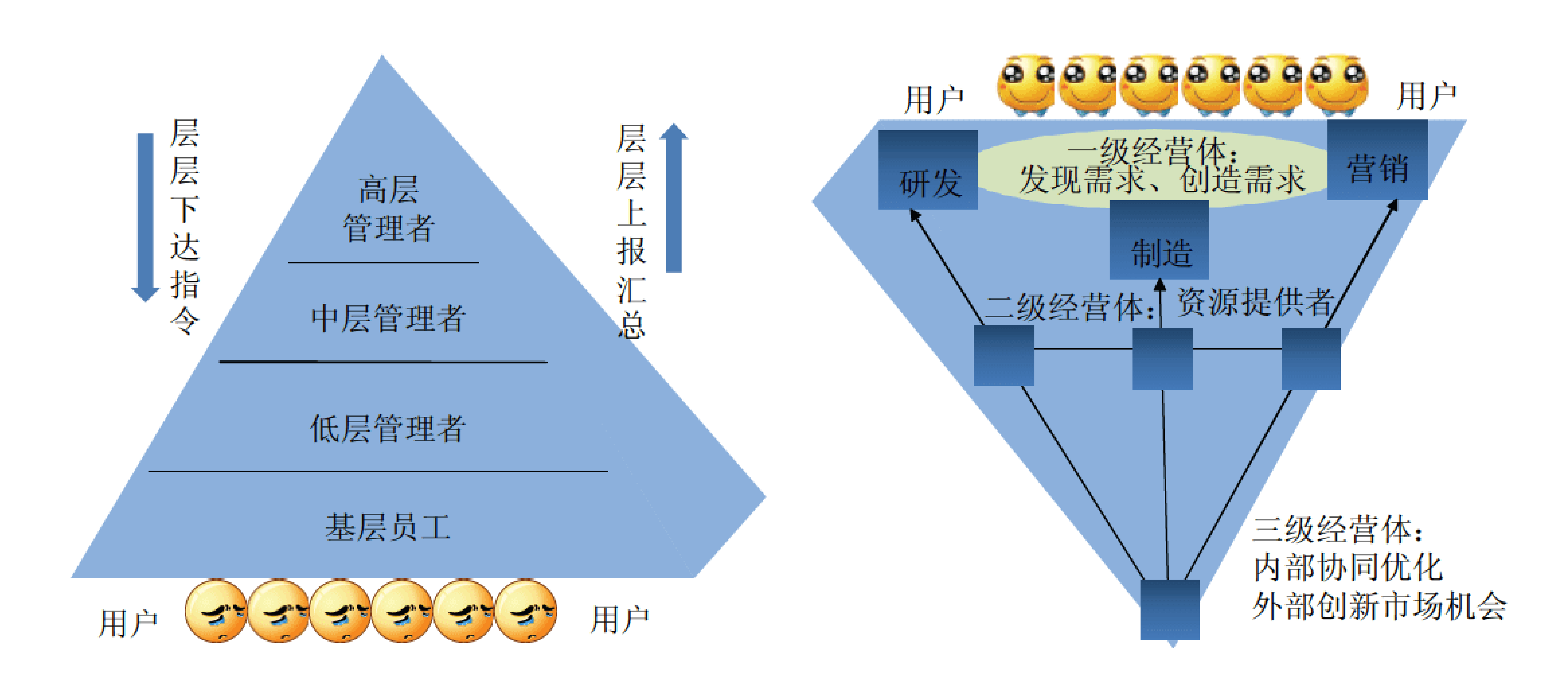


图 8-4　传统的“正三角”组织结构　　　　图 8-5　“倒三角”的组织结构

第四节　企业领导人要为组织变化做好准备

提要：

一切关乎思维。首要任务是突破窠臼，建立大数据思维。没有大数据思维一切无从谈起。

企业中组织变革常常是反应式的，在企业业绩下降、员工士气低落、指挥不灵等等征兆下，才会想到要改变。具有预见性的领导，不等这些问题出现，就会主动做好变革的准备。大数据时代，一切的变化都在加速。企业的领导应该更有前瞻性，看到大数据带来的深刻改变，面向未来，为组织变化做好准备。

唤起员工的觉醒

组织变革首先会遇到的阻力是人的惰性。多数人喜欢维持现状，喜欢过去的方式，会对变革产生巨大的抗拒力量。企业领导人面对巨大的抗拒要做好准备，有意识地创造出使员工觉醒的环境。

员工的觉醒一般是从领导者直接以改变现状的方式震撼组织成员开始的。面对现状的改变，组织成员会开始担心："那我该怎么办？"

第一，充分展现大数据，建立足够的紧迫感，以克服员工的惰性。要让员工认识到，若不拥抱大数据将是一件非常危险的事情，企业就会丧失客户，管理者就会丧失竞争力，基层员工也会丧失工作机会。领导者应该在企业内部坦诚分析一些容易使人们产生危机感的事件，如竞争对手通过大数据计划，运营效率、客户规模、客户重复购买率、客单均价等都在上升，而我们的客户正在向竞争对手转移。大数据被应用的越多，这样的事例也会越多，行业内外都能找到足够多的震撼员工的事件。外部的专家也能帮助企业认识到这样的危险，让员工感到危机逼近。

第二，建立强有力的指挥组织。除了传达组织急迫需要变革的信息外，领导者还要把一些人聚集起来，成立强有力的变革小组。小组里的成员首先要了解、支持、能够运用大数据，其次从职位、内部影响力及所拥有的信息与专业能力上都应该能够影响到变革的关键部门和环节，他们不仅要有坚定的信念，而且愿意全力投入改革行动，以追求卓越的绩效。

培养和招聘合适的人才

企业要有效获取并运用大数据，应该具备三种大数据人才：能够实施大数据 IT 基础设施的人才，能够对大数据深度分析的人才，以及知道怎么运用大数据分析结果的经理和经营分析师[①]。

能够实施大数据 IT 基础设施的人才，要能够合理评价企业原有的 IT 设施，评估原有设施在实现获取、存储、聚合、分析大数据等方面与目标的差距，在既有存量资产基础上测算所需的新的硬件、软件以及专业服务的投资。

能够对大数据深度分析的人才，要运用专业知识把系统中原始的数据化为有用的信息，他们是企业所倚重的重心。鉴于这类高端人才比较紧缺，企业应该积极招

① 参见麦肯锡，《Big data: The next frontier for innovation, competition, and productivity》，2011 年。

聘。如果需要建立一个团队，早期招聘的人才将至关重要，因为他们的水平很可能就决定了这个团队的最高水平，毕竟谁也不愿意被迅速替代。

知道怎么运用大数据分析结果的经理和经营分析师，这类人是大数据能够发挥作用的关键。他们需要经过必要的培训来获得基本的能力。企业也应该提供机会、传递压力，让经理、经营分析师与大数据分析人才充分合作，在解决问题中不断提升自我。

展望蓝图，获得支持

组织变革的过程一般是痛苦的，很多企业因此半途而废，或者偏离了以前的方向。这其中有机构的设置问题，有思想意识问题，有利益冲突问题，有业务能力问题，也有相互信任的问题……作为企业家，要善于勾勒组织的未来蓝图，为实现蓝图建立足够的共识。

推动组织变革需要将远景变得触手可及。优秀的企业家不仅具有优秀的战略能力，还能把战略演绎出坚实的逻辑，并与企业的组织能力建立关联。

2012 年，腾讯一改沿袭了 7 年的组织架构，从原有的业务部门制（Business Units，BUs）转向事业群制（Business Groups，BGs），把现有业务重新划成企业发展事业群（CBG）、互动娱乐事业群（IEG）、社交网络事业群（SNG）、网络媒体事业群（OMG）、移动互联事业群（MIG），并成立腾讯电商控股企业（ECC）独立运营电子商务业务。马化腾曾在一封内部邮件中谈到调整背后的思考。

“聚焦用户、拥抱趋势。在互联网行业，谁能把握行业趋势，最好地满足用户内在的需求，谁就可以得到用户的垂青，这个是我们行业的生存法则。今天，互联网不但已经从方方面面融入了全球 20 亿人每天的生活，也开始融入各行各业。在这个新的时代里面，用户新需求、新技术、新业务模式层出不穷，市场瞬息万变。与此同时，经过了 7 年的发展，企业的人数也超过了 2 万人，各个 BU 虽然也不断与时共进，但由于架构的限制，已经不能完全满足用户层出不穷的新需求了。所以在这个时候，我们必须要聚焦用户、顺势而变，从用户需求的角度、从产业发展的角度

重新调整我们的组织架构。”

“打造平台、产业共赢。腾讯的立业之本是我们的 IM 平台，过去的组织结构都是从这个平台上‘长’出来的，虽然枝叉变得越来越多并且落地生根，这还只是一棵树。面向未来，我们必须要向互联网更高的境界迈进。我们需要去构建一个生态系统，与合作伙伴一起培育一片森林。以前，我们主要是在企业层面思考，将来要多从产业层面思考。通过这次架构调整，企业在业务方面对各个业务群的期许是必须进一步开放思维，要有所为有所不为。一方面，在各个专业领域深耕细作，打造用户平台；但另一方面，也要培育产业让合作伙伴更好地找到共赢点。”

“大平台优势、小企业精神。2005 年进行组织架构调整的时候，全企业还只有 2000 多人，7 年来经过快速的发展，腾讯的人员规模已经是当年的 7 倍……到底我们如何能够克服大企业病，打造一个世界级的互联网企业？我们需要从‘大’变‘小’。这次调整的基本出发点是按照各个业务的属性，形成一系列更专注的事业群，减少不必要的重叠，在事业群内能充分发挥‘小企业’的精神……同时，各事业群之间可以共享基础服务平台以及创造对用户有价值的整合服务，力求在‘一个腾讯’的大平台下充分发挥整合优势。”

“推动组织变革要建立组织变革的势能。《易经》‘革’卦曰：‘已日乃孚……’，已日是收获之日‘庚日’的前一天，意思是在势能达到巅峰前顺势变革，有助于奠定基础，顺利成长。在企业建立变革的势能，一方面要让变革变得紧迫，使大家认识到在大数据的浪潮中不变则亡；另一方面要为变革做足准备，在思想意识、人才、管理能力三个层做次充分准备。”

1995 年以前，随着企业规模膨胀和人员增加，华为的部门结构日益复杂，管理难度几乎每年增加一倍，沟通难、执行难严重阻碍企业成长。1995 年 9 月，任正非亲自发起“华为兴亡，我的责任”的企业文化大讨论；1996 年 3 月，任正非邀请中国人民大学的专家起草《华为基本法》，不仅提出建设世界级企业的目标，还提出实现这一目标的路径、需要遵守的规则，后来又做了人力资源制度调整，建立规范的任职资格分级制度，并逐步完善了激励制度改革；从 1998 年开始，华为充分

积累了势能，开始建立全面国际化的组织结构，为后来的快速发展奠定了基础。

用里程碑指引未来的方向

变革需要时间。在走向未来的路上，需要一个个看得见的里程碑指引方向。如果没有这些，参与者很容易迷路，就可能失去改革的动力。换言之，如果企业只是花费大量的金钱建立了大数据系统，但是一两年内，在预测客户需求、客户服务、销售产品、产品制造、产品研发等方面都没有看到具体的应用成果，大多数参与者会感到不耐烦，整个改革方案也就无以为继，会有许多人不得不放弃改革，甚至加入反对变革的队伍中。

建立了一个个坚实的里程碑，也不能过早地宣布变革成功。过早宣布成功会减弱变革的动力，阻碍变革。既然成功了，一些改革还未到位的部分就会懈怠下去，甚至重新抬头，吞噬已经取得的部分成果。在众多的企业变革案例中，太早庆祝胜利，被抗拒者看成是改革推动者的妥协，他们赞扬推动者的改革成果，同时力劝推动者见好就收。

面对不断演进的大数据，企业领导人应该乘胜追击，推动更多的变革行动。每到达一个里程碑，应以此建立员工的信心，从而组织起更多的资源拓展更大的发展空间。

推动组织变革要刚柔并济，做好变革的过程管理。变革组织必然会损害一部分人的利益，特别是那些曾经为企业做出特别贡献的人。企业家应坚定地以企业的长远目标为出发点，通过再培训、换岗、示范等方法，完善变革过程的管理，保证推行力度，才能实现组织的有效变革。

大数据是推动组织变革的动因，但并不能必然带来健康的组织变革。只有创新企业管理，善于运用大数据，才会让企业的组织变革走向成功。大数据浪潮也终将埋葬企业传统的组织形态，不断发育出新的组织形态。我们站在浪潮之巅是，与之共舞，还是被它淹没？

第二部分

数据科学

大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑战。它将对现有的科研和教学体制、科学与产业之间的关系、科学与社会之间的关系带来大幅度的变革。用数据来研究科学，科学地研究数据。数据科学地兴起和发展，将深刻改变人类探索世界的思维和方法。

导读：

1. 数据科学主要包括两个方面：用数据的方法来研究科学和用科学的方法来研究数据。前者包括象生物信息学、天体信息学、数字地球等领域。后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分。但只有把它们有机地放在一起，才能形成整个数据科学的全貌。
2. 在数据科学领域里工作的人才需要具备两方面的素质：一是概念性的，主要是对模型的理解和运用；二是实践性的，主要是处理实际数据的能力。培养这样的人才，需要数学、统计和计算机科学等学科之间的密切合作，同时也需要和产业界或其他拥有数据的部门之间的合作。目前还没有任何一所高校具有这样的平台。

第九章

数据科学

数据科学将逐渐达到与其他自然科学分庭抗礼的地位。

——笔者

大数据时代在科学领域里的表现是数据科学的兴起。常常听到有人问：多大才算是“大数据”？“大数据”和“海量数据”有什么区别？其实根本没有必要为“大数据”这个名词的确切含义而纠结。“大数据”是一个热点名词，它代表的是一种潮流、一个时代，它可以有多方面的含义。“海量数据”是一个技术名词，它强调数据量之大。而数据科学则是一门新兴的学科。

为什么要强调数据科学？它和已有的信息科学、统计学、机器学习等学科有什么不方一样？

第一节　数据科学的基本内容

作为一门学科，数据科学所依赖的两个因素是数据的广泛性和多样性，以及数据研究的共性。现代社会的各行各业都充满了数据，而且这些数据也是多种多样的，不仅包括传统的结构型数据，也包括网页、文本、图像、视频、语音等非结构型数据。正如后面将要讨论到的，数据分析本质上都是在解反问题，而且是解随机模型的反问题。所以对它们的研究有着很多的共性。例如，自然语言处理和生物大分子模型里都用到隐式马氏过程和动态规划方法，其最根本的原因是它们处理的都是一维的随机信号。再如图像处理和统计学中都用到的正则化方法，也是处理反问题的数学模型中最常用的一种手段。所以用于图像处理的算法和用于压缩感知的算法有着许多共同之处。这在新加坡国立大学沈佐伟教授的工作中就可以很明显地看出来。

除了新兴的学科如计算广告学之外，数据科学主要包括两个方面：用数据的方法来研究科学和用科学的方法来研究数据。前者包括生物信息学、天体信息学、数字地球等领域，后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分，但只有把它们有机地放在一起，才能形成整个数据科学的全貌。

用数据的方法来研究科学，最典型的例子是开普勒关于行星运动的三大定律。

开普勒的三大定律是根据他的前任，一位名叫第谷的天文学家留给他的观察数据总结出来的。表 9-1 是一个典型的例子，这里列出的数据是行星绕太阳一周所需要的时间（以年为单位）和行星离太阳的平均距离（以地球与太阳的平均距离为单位）。从这组数据可以看出，行星绕太阳运行的周期的二次方和行星离太阳的平均距离的三次方成正比。这就是开普勒的第三定律。

表 9-1 太阳系八大行星绕太阳运动的数据

行星	周期/年	平均距离	周期 ^{2}/距离 ^{3}
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

开普勒虽然总结出行星运动的三大定律，但他并不理解其内涵。牛顿则不然，牛顿用他的第二定律和万有引力定律把行星运动归结成一个纯粹的数学问题，即一个常微分方程组。如果忽略行星之间的相互作用，那么这就成了一个两体问题。因此很容易求出这个常微分方程组的解，并由此推出开普勒的三大定律。

牛顿运用的是寻求基本原理的方法，远比开普勒的方法深刻。牛顿不仅知其然，而且知其所以然。所以牛顿开创的寻求基本原理的方法成了科学研究的首选模式。这种方法在 20 世纪初期达到了顶峰：在它的指导下，物理学家们发现了量子力学。从原则上来讲，日常生活中的自然现象都可以从量子力学的角度来解释。量子力学提供了研究化学、材料科学、工程科学、生命科学等几乎所有自然和工程学科的基本原理。这应该说是很成功的，但事情远非这么简单。正如狄拉克指出的那样，如果以量子力学的基本原理为出发点去解决这些问题，那么其中的数学问题太难了。

所以如果要想有进展，还是必须做妥协，也就是说要对基本原理作近似。

再举另外一个例子，表 9-2 中形象地描述了一组人类基因组的 SNP（Single Nucleotide Polymorphism）数据。一组研究人员在全世界挑选出 1064 个志愿者，并把他们的 SNP 数据数字化，也就是把每个位置上可能出现的 10 种碱基对用数字来代表，对这组数据作主组分分析，就可以得到图 9-1 中的结果。图中横轴和纵轴代表的是第一和第二奇异值所对应的特征向量，这些向量一共有 1064 个分量，对应 1064 个志愿者。值得注意的是，这组点的颜色所代表的意义。可以看出，人类进化的过程可以从这组数据中通过最常见的统计分析的方法，即主组分分析方法而展示出来。

主组分分析是一种最简单的数据分析方法。它的做法是对数据的协方差矩阵作对角分解。

表 9-2 SNP 数据的示意图

	SNP1	SNP2	…	SNPm
志愿者 1	0	1	…	0
志愿者 2	0	2	…	1
志愿者 3				
⋮	⋮	⋮	⋮	⋮
志愿者 n	1	9	…	1

注：n=1064，m=644258，0，1，…，9 分别代表碱基对是 AA，AC，CC，…[①]

这样的问题，如果采用从基本原理出发的牛顿模式，那么基本上是没法解决的，而基于数据的开普勒模式则是行之有效的。尽管牛顿模式很深刻，但对复杂的问题，开普勒模式往往更有效。开普勒模式最成功的例子是生物信息学和人类基因组工程。正是因为它们的成功，材料基因组工程等类似的项目也被提上了议事日程。同样，

①参见：Jun Z.Li et al,"Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation"，Science,22,Feburary，2008。

天体信息学、计算社会学等等也成了热门学科。这些都是用数据的方法来研究科学问题的例子。图像处理是另外一个典型的例子。图像处理是否成功是由人的视觉系统决定的，所以要从根本上解决图像处理的问题，就需要从理解人的视觉系统着手，并了解不同质量的图像对人的视觉系统产生什么样的影响。这样的理解当然很深刻，而且也许是大家最终所需要的。但从目前来看，它过于困难也过于复杂。解决很多实际问题时并不需要它，而是用一些更为简单的数学模型就足够了。

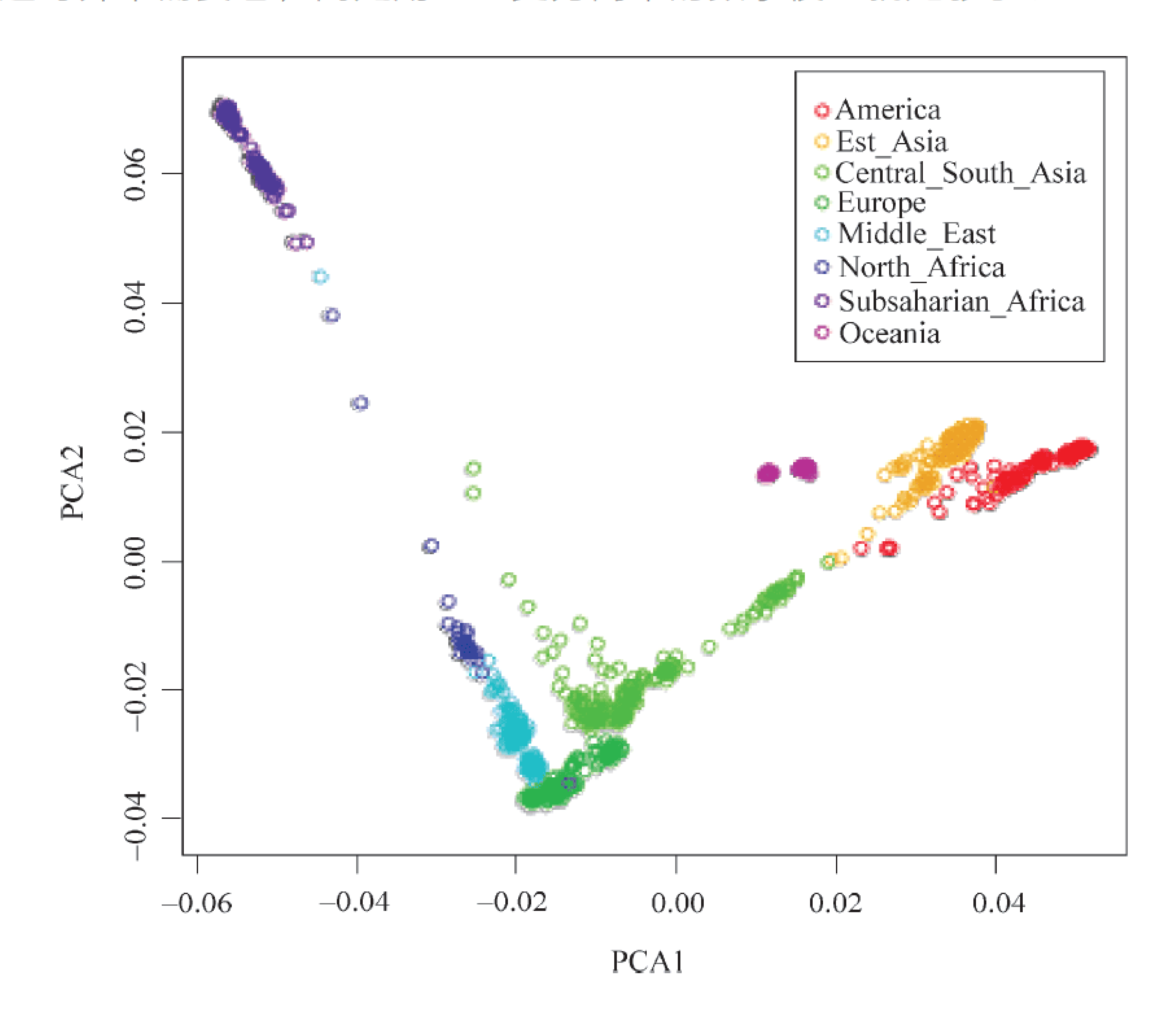


图 9-1　对 SNP 数据作主组分分析的结果展示了人类进化的过程[①]。

用数据的方法来研究科学问题，并不意味着就不需要模型了。只是模型的出发点不一样，不是从基本原理的角度去找模型。就拿图像处理的例子来说，基于基本原理的模型需要描述人的视觉系统以及它与图像之间的关系，而通常的方法则可以

① 这里横轴和纵轴分别表示最大奇异值和第二大奇异值所对应的特征向量。此结果系姚远等根据 Jun Z. Li 等文章中的结果重新制作的。

是基于更为简单的数学模型，如函数逼近的模型。

怎样用科学的方法来研究数据？这包括以下几个方面的内容：数据的获取、存储和数据的分析。下面将主要讨论数据的分析。

数据分析的中心问题

比较常见的数据有以下几类：

1. 表格。这是最为经典的数据。

2. 点集（point cloud）。很多数据都可以看成是某种空间中的一堆点。

3. 时间序列。文本、通话、DNA 序列等都可以看成是时间序列。它们也是一个变量（通常可以看成是时间）的函数。

4. 图像。可以看成是两个变量的函数。

5. 视频。时间和空间坐标的函数。

6. 网页、报纸等。虽然网页或报纸上的每篇文章都可以看成是时间序列，但整个网页或报纸又具有空间结构。

7. 网络数据。

还可以考虑更高层次的数据，如图像集、时间序列集、表格序列等等。

数据分析的基本假设就是观察到的数据都是由背后的一个模型产生的。数据分析的基本问题就是找出这个模型。由于数据采集过程中不可避免地会引入噪声，通常这些模型都是随机模型，见表 9-3。

表 9-3　数据类型与模型

数据类型	模　　型
点集	概率分布
时间序列	随机过程（如隐式马氏过程等）
图像	随机场（如吉布斯随机场）
网络	图模型、贝叶斯模型

当然，在大部分情况下，整个模型并不令人感兴趣，而找到模型的一部分内容

是需要关注的东西，例如：

1. 相关性。判断两组数据是不是相关的。

2. 排序。如对网页作排序。

3. 分类、聚类。把数据分成几类。

很多情况下，还需要对随机模型作近似。最常见的是把随机模型近似为确定模型，所有的回归模型都采用了这样的近似，基于变分原理的图像处理模型也采用了同样的近似。另一类方法是对其分布作近似，如假设概率密度是正态分布，或假设时间序列是马尔可夫链等等。

分析数据的第一步是赋予数据一定的数学结构，这种结构包括：

1. 度量结构。在数据集上引进度量，也就是距离，使之成为一个度量空间，余弦距离函数。

2. 网络结构。有些数据本身就具有网络结构，如社交网络。有些数据本身没有网络结构，但可以附加上一个网络结构。例如，度量空间的点集，可以根据点与点之间的距离来决定是否把两个点连接起来，这样就得到一个网络结构。

3. 代数结构。例如，可以把数据看成是向量或矩阵，或更高阶的张量。有些数据集具有隐含的对称性，这也可以用代数的方法表达出来。

在这基础上，可以问更进一步的问题，例如：

1. 拓扑结构。从不同的尺度去看数据集，得到的拓扑结构可能是不一样的。最著名的例子是 3×3 的自然图像数据集里面隐含着一个二维的克莱因瓶[①]。

2. 函数结构。尤其对点集而言，寻找其中的函数结构是统计学的基本问题。这里的函数结构包括：线性函数，用于线性回归；分片常数，用于聚类或分类；分片多项式，如样条函数；其他函数，如小波展开等。

---

① 参见： Robert Ghrist, BARCODES: THE PERSISTENT TOPOLOGY OF DATA, BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY， Volume 45, Number 1, January 2008, Pages 61 - 75.

数据分析的主要困难

人们碰到的数据通常有这样几个特点：一是数据量大，大家只要想一想万维网上有多少网页，这些网页上有多少数据，就可以对现在碰到的数据量之大有点感觉了；二是维数高，前面提到的 SNP 数据是 64 万维的；三是类型复杂，如这些数据可以是网页或报纸，也可以是图像、视频；四是噪声大。

这里面最核心的困难是维数高。维数高带来的是维数诅咒（curse of dimension）：模型的复杂度和计算量随着维数的增加而指数增长。例如，非参数化的模型中参数的个数会随着维数的增加而指数增长。

怎样克服维数高带来的困难？通常有两类方法：一类方法就是将数学模型限制在一个极小的特殊类里面，如线性模型，假设概率密度遵循正态分布、假设观测到的时间序列是隐式马氏过程等；另一类方法是利用数据可能有的特殊结构，如稀疏性、低维或低秩、光滑性等等，这些特性可以通过对模型作适当的正则化而实现。当然，降维方法也是主要方法之一。

总而言之，数据分析本质上是一个解反问题。因此，处理反问题的许多想法，如正则化，其在数据分析中扮演了很重要的角色。这也正是统计学与统计力学的不同之处。统计力学处理的是正问题，统计学处理的是反问题。

算法的重要性

跟模型相辅相成的是算法以及这些算法在计算机上的实现。特别是在数据量很大的情况下，算法的重要性就显得尤为突出。

从算法的角度来看，处理大数据主要有两条思路：

一是降低算法的复杂度，即计算量。通常要求算法的计算量是线性标度的，也就是说计算量跟数据量成线性关系。但很多关键的算法，尤其是优化方法，还达不到这个要求。对特别大的数据集，如万维网上的数据或社交网络数据，许多人希望

能有次线性标度的算法，也就是说计算量远小于数据量。这就要求采用抽样的方法。但怎样对这样的数据进行抽样，如对社交网络进行抽样，仍还是一个未解决的问题。

二是云计算，或并行计算。它的基本想法是把一个大问题分解成很多小问题，然后分而治之。著名的 MapReduce 软件就是一个这样的例子。

下面举几个典型的算法方面的例子。这些例子来自于 2006 年 IEEE 国际数据挖掘会议所选举出来的数据挖掘领域中的 10 个最重要的算法。

1. k-平均（k-means）方法。这是对数据作聚类的最简单有效的方法。

2. 支持向量机：一种基于变分（或优化）模型的分类算法。

3. 期望最大化（EM）算法。这个算法的应用很广，典型的应用是基于极大似然方法（maximum likelihood）的参数估计。

4. 谷歌的网页排序算法，PageRank。它的基本想法：网页的排序应该是由网页在整个互联网中的重要性决定的。从而把排序问题转换成一个矩阵的特征值问题。

5. 贝叶斯方法。这是概率模型中最一般的迭代法框架之一。它告诉人们怎样从一个先验的概率密度模型，结合已知的数据来得到一个后验的概率密度模型。

6. k-最近邻域方法。用邻域的信息来作分类。跟支持向量机相比，这种方法侧重局部的信息。支持向量机则更侧重整体的趋势。

7. AdaBoost。这个方法通过变换权重，重新运用数据的办法，把一个弱分类器变成一个强分类器。

8. 其他的方法。例如，决策树方法和用于市场分析的 Apriori 算法，以及用于推荐系统的合作过滤方法等。

就现阶段而言，对算法的研究被分散在两个基本不相往来的领域里：计算数学和计算机科学。计算数学研究的算法基本上是针对像函数这样的连续结构，其主要的应用对象是微分方程等。计算机科学处理的主要是离散结构，如网络。而数据的特点介于两者之间，数据本身当然是离散的，但往往数据的背后有一个连续的模型。所以要发展针对数据的算法，就必须把计算数学和计算机科学研究的算法有效地结合起来。

第二节　对学科发展的影响

回到本章的主题，数据科学对学科发展提供了前所未有的机遇和挑战。要充分利用好这个机会，就必须建立起一套新的科学和教育体系。在大学的层面，要赋予数据科学应有的地位，建立起跨学科、全方位的数据科学研究平台；进一步完善和企业合作创新的机制；培养适应学术界和企业界需求的数据科学人才。

数据科学也将对许多传统学科的发展带来极大的影响。首先是对数学，数学的发展主要来自两个方面的推动力：一是来自数学内部，学科自身的完善带来的推动；二是来自外部，由其他学科、社会或工业发展的需要而带来的推动。就目前的现状而言，第一方面的推动力对数学的影响要远远超过第二方面的推动力。这样造成的结果是：一方面，数学作为一门学科，其重要性已经得到广泛的认可；而另一方面，数学家作为一个群体，其对社会和科学整体发展的影响却难以得到承认。在很多学校以及在整个科学界，数学家这个群体正显得越来越孤立。这就是为什么数学家们经常发现自己处在一个很尴尬的位置。这是一件极为不幸的事情，它不仅大大影响了数学的发展，更是影响其他学科、技术乃至社会的发展。事实上，至少在理论研究方面，很多学科的瓶颈问题都是数学问题。这在近一百年前狄拉克就已经指出来了。所以在很多学科里，人们看见的都是非数学出身的科学家在进行数学方面的研究。

数学家们为什么不擅于帮助解决其他学科的问题呢？在自然科学领域，有一个基本的原因，那就是要解决自然科学的问题首先要有基本原理，也就是通常所说的模型。人们把它们叫做数学模型。但实际上这些模型都是来自于物理学的基本原理。对数学家们来说，这是一个基本障碍。

数据科学不一样，如前所述，数据科学的基本原理本身就来自于数学。所以数据科学在数学和实际应用之间建立起了一个直接的桥梁。而这些实际应用正是来自

于如信息服务等现代产业中最为活跃的一部分。这对数学来说，实在是一个千载难逢的机会。

不仅如此，数据的分析几乎涉及到了现代数学的所有分支，甚至于像表示论这样的极其抽象的分支在数据的领域也有其发挥作用的余地。所以数据科学对数学的要求和推动是全面的，而不是仅仅局限在几个领域。数据应该成为数、图形和方程之外数学研究的基本对象之一。

数据科学对计算机科学的发展也会带来很大的影响。图灵奖得主 John Hopcroft 曾经 指出，在过去的几十年里，计算机科学的研究对象主要是计算机本身，包括硬件和软件。以后计算机科学的发展将主要围绕着应用展开。而从计算机科学自身来看，这些应用领域提供的主要研究对象就是数据。虽然计算机科学一贯重视数据的研究，但数据在其中的地位将会得到更进一步的加强。

再看统计学。统计学一直就是一门研究数据的学科，所以它也是数据科学最核心的部分之一。但在数据科学的框架之下，统计学的发展也会受到很大的冲击。这种冲击至少表现在两个方面。一是关于数据的模型将会跳出传统的统计模型的框架，更一般的数学概念，如拓扑、几何和随机场的概念将会在数据分析中扮演重要的角色；二是算法和计算机上的实现将成为研究的中心课题之一，这在前面已经讨论过，这里不再重复。

应该说，在很长的一段时间里，统计学这门学科没有受到足够的重视。普林斯顿大学还取消了统计系。近年来，学术界和应用领域都已经逐渐地认识到统计的重要性。许多学校都有计划要发展统计学，但苦于难以吸引到高质量的统计人才而迟迟没有开展。如果把视野拓宽一点就会发现，发展数据科学则是更加有利的做法，因为它既更加适应未来的需要，又能尽快地把应用数学、计算数学和计算机科学等学科中的有生力量调动起来以开展工作，如图 9-2 所示。

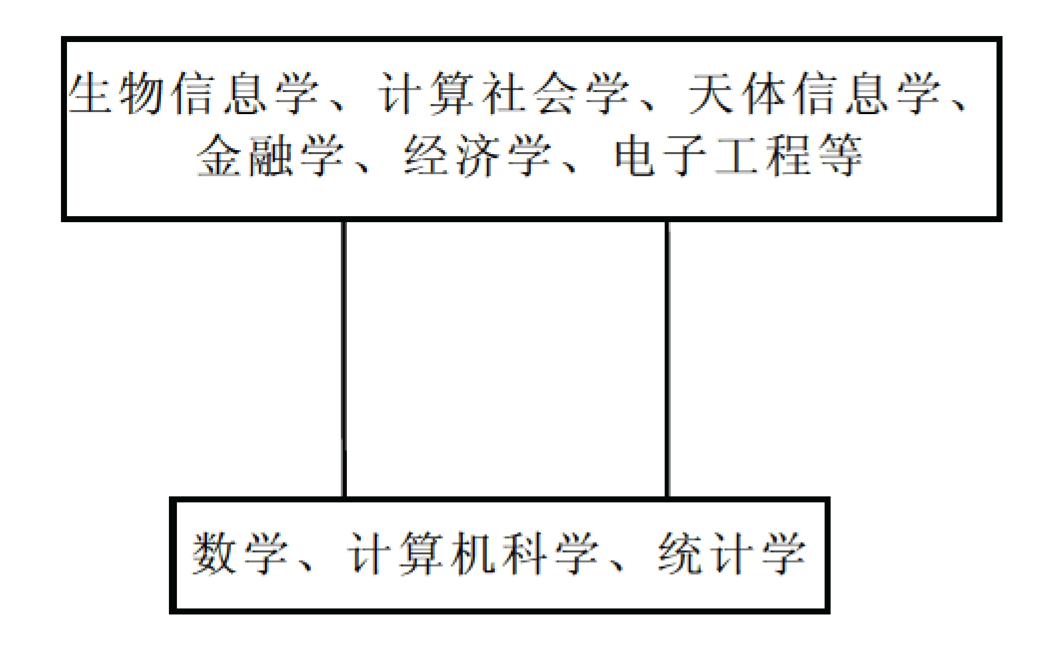


图 9-2　数据科学的基础性作用

对传统学科的冲击

这里举两个例子。第一个例子是社会学。作为社会科学的一个分支，社会学一直是一门基于数据的学科。大到国家和社会层面的数据，小到家庭和个人的数据，这些是社会学研究的基本资料。从这个角度来看，社会学和数据之间的关系不是什么新的现象。但即便以此，数据科学的兴起仍然对社会学的研究有着巨大的影响。这至少表现在以下几个方面。

一是社交网络的产生和网络科学的研究为社会学带来了一个新的研究层面，即介观层面。这不仅给社会学提供了新的研究方向，而且也给社会学的研究提供了新的实用价值，如信息传播、广告投放、热点分析等。

二是使社会学的研究进一步量化、去经验化。在过去很长的时间里，由于数据的稀缺，社会学在很大程度上是一门经验科学。大量数据资源的获取为社会学的更进一步量化提供了可靠的途径。

三是更多更加严密和系统的科学方法被引进到社会学的研究中，如数据采集的方法。北京大学中国社会调查中心所开展的家庭访问调查就是一个很好的例子。他们不但注重调查中问答的结果，同时也记录了调查过程的数据。这样严密的科学方法一定会给社会学的研究带来极大的影响。

在人们眼里，社会学往往不是一门技术型的或实用型的学科。但随着社会学的进一步量化，人们对社会学的看法将会发生很大的变化。在不远的未来，社会学的

研究将对产品推销、信息传播和舆情预警等实用领域产生深刻的影响。

第二个例子是语言学。跟社会学一样，语言学在历史上也是一个离实用技术比较远的学科。但近年来蓬勃发展起来的机器翻译、自然语言处理、语言识别、文本分析等技术给语言学的实际应用提供了一个绝好的机会。但值得注意的是，在所有这些领域，基于概率模型的处理方法的有效性远远超过了基于文法的处理方法的有效性。这对传统的语言学来说，不能不说是一个非常令人失望的结果。

在麻省理工学院成立 150 周年的一个纪念会上，当代语言学的奠基人乔姆斯基教授针对这一问题提出了他的看法。他认为概率模型的成功是有限的，而且其成功只是仅仅局限于逼近未被分析的数据这一方面。他的言下之意是说概率模型只是技术上的成功，不能算作是传统科学意义上的成功，因为它没有给传统的语言学问题如文法问题，带来新的认识。应该说，这种看法是比较保守的，按照这种逻辑，生物信息学也只是工程上的成功，不是科学意义上的成功。按照前文的说法，自然语言的概率模型可以看成是一种开普勒模式的做法，而乔姆斯基只认可牛顿模式。科学发展的历史已经告诉人们，这两种模式都十分重要。而具体到语言学来说，承认并认真应对概率模型的成功才是真正可取的方法。

新学科的诞生：计算广告学

广告有着十分悠久的历史，但它一直都很难算得上是一门科学。尤其是在中国，由于管理上的漏洞，最典型的广告就是在媒体上，特别是在电视上，由各种各样的明星说上几句不负责任的话。近年来，由于雅虎、谷歌等搜索引擎选择商业广告作为其主要盈利模式，一门新的学科——计算广告学，由此诞生。

计算广告学所处理的主要问题是怎样有针对性地投放广告。互联网上的广告有两个最基本的指标：点击率和转换率。点击率是广告被点击的概率。转换率是广告被点击以后引起商品成交的概率。由于后者更难估计，所以互联网上的广告往往以点击率作为主要指标。这就需要根据用户提供的信息，如其所输入的关键词，预测

不同广告的点击率。这是计算广告学的一个基本问题，解决这个问题的主要想法就是构造一个 utility 函数来估计用户对不同广告感兴趣的程度。

目前斯坦福大学、加州大学伯克利分校等重要学校都已开设了计算广告学这门课。美国国家基金委所属的几个数学研究所之一，地处北卡罗来纳州的统计与应用数学研究所也针对计算广告学举办了专题研讨会。

第三节　科学能从谷歌那儿学到什么？

“科学能从谷歌那儿学到什么？”是 2008 年美国《连线》杂志（Wired Magzine）主编安德森在他的一篇评论文章（The end of theory: The data deluge makes the scientific method obsolete, Wired Magazine, 06.23.08）结尾时的问话。的确，谷歌不仅仅是信息产业界成功的典范，同时还是数据科学领域的先锋和开拓者。谷歌的成长史是一部创新和开拓的历史。

谷歌的起步是源于网页搜索排序的新概念和算法开发。谷歌之前已经有了其他的搜索引擎，最著名的是雅虎。但所有这些引擎都没有解决好对搜索结果作排序的问题。佩奇和布林的想法是把网络的结构利用起来。事实上，每个网页都是互联网上的一个节点，它们不是孤立的，不同的网页之间通过超链接联系在一起。如果一个网页有很多超链接指向它，就说明它具有权威性，应该排在前面。怎样给网页的权威性一个定量的刻划呢？设想一个醉汉在互联网上作随机游动，他访问的最多的网页就最具有权威性。这样就可以把网页排序的问题描述成为一个由互联网结构而派生出来的马氏链的不变测度的问题，也就是一个转移矩阵的特征值问题。这就是佩奇关于网页排序的基本想法。通过这种想法，佩奇和布林大大提高了互联网搜索结果的质量。

谷歌也是第一个将云计算由概念变为现实的企业。不言而喻，谷歌从一开始就需要处理大量的网页。它最初开发云计算的目的是建立一个能把大量的廉价服务器

集合在一起，以完成大型计算和存储的功能平台。这个平台必须是可扩展的、并行的，并且允许其中一些服务器出现故障。为了达到这一目的，谷歌开发了一系列的新技术和新的数据存储模式，其中包括谷歌文件系统（Google File System）、MapReduce 等。这些新概念和新技术已成为大数据处理的标准方法。与此同时，谷歌也建立起了面向未来的数据中心和云计算平台。这些基础设施使得谷歌在信息服务产业高居于一个得天独厚的位置。

谷歌之所以能做到这些，最根本的一点是它高瞻远瞩的眼光和宽广的胸怀。谷歌创始人佩奇和布林认识到，谷歌的根本利益在于互联网能否成为普通大众生活中必不可少的工具。做好了这一点，谷歌的商业利益就自然而然地来了。为了做到这一点，谷歌坚持了由雅虎开创的互联网免费的原则。这个原则对互联网的普及起到了最为关键的作用。

事实上，谷歌的商业模式也是可圈可点的。它的盈利是靠互联网广告，而不是靠对用户的收费。在谷歌之前，Overture 公司就已经在开展互联网广告业务，但谷歌把互联网广告推到了更高的层次。谷歌开发的 Adwords 系统是计算广告学最早的实践典范。

互联网是一个极大的资源，一个由全世界的亿万网民共同构建的资源。而谷歌这样的公司，通过构建一系列新的概念和技术平台，十分有效地把这些资源变成了他们自己的资源。而在此同时，又给全世界的网民提供了十分有益的服务。谷歌的例子是创新和产业发展密切结合、相互推动最成功的例子。

第四节　数据科学的教育体系

在数据科学领域里工作的人才需要具备两方面的素质：一是概念性的，主要是对模型的理解和运用；二是实践性的，主要是处理实际数据的能力。培养这样的人才需要数学、统计学和计算机科学等学科之间的密切合作，同时也需要和产业界或

其他拥有数据的部门之间的合作。目前还没有任何一所高校具有这样的平台。

数据科学的教育体系应该包括以下几方面的内容：

1. 数学的基础知识。除了微积分、线性代数和概率论这三大基础中的基础以外，还需要随机过程、函数逼近论、图论、拓扑学、几何、变分法、群论等方面的基础知识。目前，可能还不是所有人都能看到这些内容跟数据的直接关系。但随着数据科学的不断深入发展，它们的作用会越来越明显。这些内容也不需要一门一门地教，数学系应该开出一些新的“高等数学”课程来覆盖这些方面的内容。

2. 计算机科学的基本知识。例如，计算机语言、数据库、数据结构、可视化技术等。

3. 算法方面的基本知识。例如，数值代数、函数逼近、优化、蒙特卡洛方法、网络算法、计算几何等等。

4. 数据的模型。例如，回归、分类、聚类、参数估计等。

5. 专业课程。例如，图像处理、时间序列分析、视频处理、自然语言处理、文本处理、语言识别、图像识别。推荐系统等等。

6. 其他专业课。例如，生物信息学、天体信息学、金融数据分析等等。

这里 1～4 属于基础课，5～6 属于专业课。专业课的设置还可以跟企业界合作，以满足不断变化着的实际需求。与企业界的合作也更有利于向企业界输送合适的人才。

结束语

大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑战。它不仅将给现有的科研和教学体制带来大幅度的变革，也会给科学与产业之间的关系、科学与社会之间的关系带来大幅度的变革。总结一下，大数据的影响将主要来自以下几个方面：

1. 数据科学将成为科研体系中的重要部分，并逐渐达到与包括物理、化学、生

命科学等学科在内的自然科学分庭抗礼的地位。未来的科研和教育体制应该由两条主线组成：一条是以基本原理为主线。现在的物理学、化学、机械工程等学科，以及生命科学、材料科学、天体物理、地球科学等学科的大部分都是沿着这样一条主线展开的。另一条是以数据为主线。它包括统计学、数据挖掘和机器学习、生物信息学、天体信息学、以及许多社会科学的学科。它还包括一些新兴的学科，如计算广告学。数据科学的兴起，将极大地推动许多社会科学学科朝着量化的方向发展，使他们逐步由经验性的模式转变成科学性的模式。

2. 科学研究和市场、产业的联系将变得更加密切，从发现基本原理到产业化的周期将会被大大地缩短。这可以从谷歌的例子看出来。谷歌的发展，从搜索引擎的一个概念和算法上的突破到进入市场、变成产业，只经过了短短几年的时间。这样的例子在数据科学和信息产业领域并不陌生。但在传统的自然科学领域，从基本原理的突破，到技术、到产业，往往要经过一个漫长的过程。

3. 数据的主要来源之一是社会，如互联网、社交网络、公共交通、智慧城市等等。所以数据科学的研究与人们的日常生活和社会有着密切的联系，如谷歌和百度的网络搜索算法就对人们的日常生活产生了很大的影响。所以人们日常生活中的需要以及社会的需要将成为数据科学的主要问题来源之一。

4. 科学研究最重要的一环是提出前瞻性的问题。提不出问题，就只能跟在别人后面，走一条从文献到文献的路子。对我国的科技界来讲，很多学科由于来自实际应用领域的限制，提出前瞻性问题的确是件很困难的事情。但数据科学则不然，由于我国人口众多这一特殊情况，加上我国特殊的文化、文字、历史背景和社会发展的需要，在数据科学领域的很多问题自然就是前瞻性的，关键在于能否用前瞻性的方法去面对这些问题。如果做好了这一点，我国在数据科学领域就会自然而然地走到了世界的前沿。

导读：

1. Hadoop 相关技术，如同 PC 时代的 Windows 操作系统，这些技术为企业构建大数据处理平台提供了基础的系统架构，及相关的数据库、数据流等数据管理工具。
2. 大数据时代的数据挖掘技术，由于所需要处理的数据规模庞大、且价值密度低，在处理方法和逻辑上被赋予了新的含义。
3. 直觉式的市场决策已不能适应企业竞争的需求，越来越多的业务依赖于对内外部数据的深度解读。数据专家在未来企业竞争和战略发展中的作用将发挥越来越大的作用。

第十章

数据技术：当前进展及关键问题

欲工其事必先利其器。

——笔者

大数据时代，企业的核心竞争力将取决于其占有数据的规模、活性及对数据的分析和运用能力。互联网、物联网等技术在各个行业的普及，使企业内部或外部产生的数据规模急剧增加，并且数据种类繁多，企业和个人用户信息使用模式也千变万化。所有这些都对企业大规模数据的收集、存储和处理能力提出了越来越高的要求。

数据技术是为满足企业在大数据时代的数据处理需求而发展起来的数据采集、过滤、存储、变换、分析和挖掘等一系列相关工具、技术的总称。由于数据规模庞大，对实时性要求高，原有的数据采集、存储等技术已无法应对大数据时代的需求。所谓工欲善其事，必先利其器。对先进数据技术的掌握和运用能力是企业在大数据时代保持领先水平的技术基础。幸运的是，目前在数据采集、存储、分析及应用的各个层面，都有相对比较成熟的技术可供选择。其中很多甚至是能够满足企业级应用的开源软件。

这方面值得关注的是谷歌的大数据处理能力及在其基础上发展起来的 Hadoop 相关技术。如同 PC 时代的 Windows 操作系统，这些技术为企业构建大数据处理平台提供了基础的系统架构，及相关的数据库、数据流等数据管理工具。

虽然数据体量庞大是大数据时代的特点，但这并不意味着数据的含金量高，对数据的理解要求低。事实上，庞大的数据中往往掺杂着各种噪音或无效数据，其单位含金量更低。简单粗放式的数据统计和分析往往不能得到真正有价值的内容，甚至可能是相左的结论，所以需要更加有效的、精工细作模式的处理能力。这些，无论是从数据处理规模，还是从算法的健壮性等方面来看，都对相关的数据挖掘技术提出了更高层次的挑战。

本部分从 Hadoop 系统和数据挖掘技术两个角度讲述当前数据技术的进展和面临的挑战。最后介绍在大数据时代企业需要什么样的数据分析和挖掘能力，及大数据时代的弄潮儿——数据专家是如何炼成的。

第一节 大数据管理系统——Hadoop

似乎从诞生之日起，Hadoop 便与大数据有着千丝万缕的联系。Hadoop 的设计原理来源于谷歌的 GFS 和 MapReduce 模型，可以看作是后者的开源实现。由于其可以运行在对硬件配置要求低、扩展性好、容错能力强及强大的并行处理能力等特点的设备上，在多个行业得到广泛应用，成为当下大数据领域的热门技术。

那么为什么企业的数据处理需要 Hadoop 技术？相对其他系统，它能为企业带来什么样的技术优势？该系统具体包括哪些技术？首先用两个例子说明 Hadoop 在大数据存储和管理中的独特优势。

首先以 Facebook 为例说明用户行为数据的极大膨胀为数据存储带来的挑战。做为最大的社交网站，Facebook 拥有超过 5 亿的活跃用户，在 Facebook 上分享了 2400 亿张图片，仅图片存储容量就达 20PB 的规模，且仍以每天 3000 多万新增图片的规模迅速扩张[①]。此外，每月还会产生超过 250 亿条的分享内容信息，及超过 5000 亿次的页面流量记录。要存储如此大规模的数据内容，且支持内容的随机读取操作，无疑给系统的存储和处理能力带来了极大挑战。

同时，如果仅仅是静态的存储数据，对企业来说无疑是毫无价值的。Facebook 的迷人之处在于其对海量数据的快速研究应用能力，让原本杂乱无章的数据真正流动起来。从简单的统计功能，如不同页面的浏览量、用户行为数据，到诸如用户兴趣类别划分、内容推荐等复杂模型的建立，可以想象完成这么多大规模数据处理任务需要什么样的运算能力。

早期 Facebook 的数据仓库是基于 Oracle 系统实现，随着用户数据的增多，在系统可扩展性和系统性能方面都遇到了瓶颈。2008 年之后，Facebook 开始采用 Hadoop/Hive 等技术搭建其数据仓库。截至 2011 年，其数据仓库已拥有 4800 个

① Julie Bort. Facebook Stores 240 Billion Photos And Adds 350 million More A Day. Jan. 2013. http://www.businessinsider.com/facebook-stores-240-billion-photos-2013-1。

内核，每个节点可存储超过 10TB 的数据[①]。

图 10-1 所示为 Facebook 的数据仓库架构。Facebook 每天产生的所有日志数据都会存储在文件管理系统——HDFS 中。拥有如此庞大的用户群，对数据安全的要求是异常之高的。很难想象，如果用户某天突然发现其 Facebook 页面的数据丢失或出错了，用户体验和公司的声誉会受到什么样的影响。HDFS 通过采取数据冗余设计机制从而具有良好的容错特性，避免由于硬件损坏带来的数据丢失影响。同时，HDFS 良好的可扩展性，使其能够支持数据不断增长带来的挑战。采取这种架构，Scribe 组件可以支持海量用户行为日志的连续读写，实现快速的数据收集。以 Hive 为基础的数据仓储中心可以及时将收集的日志写入 HDFS，并支持各种数据分析进行的统计工作。用户加载页面时需要实时从数据库中获取相关信息，以 HBase 为核心的实时随机读写模块则可以有效实现该需求。

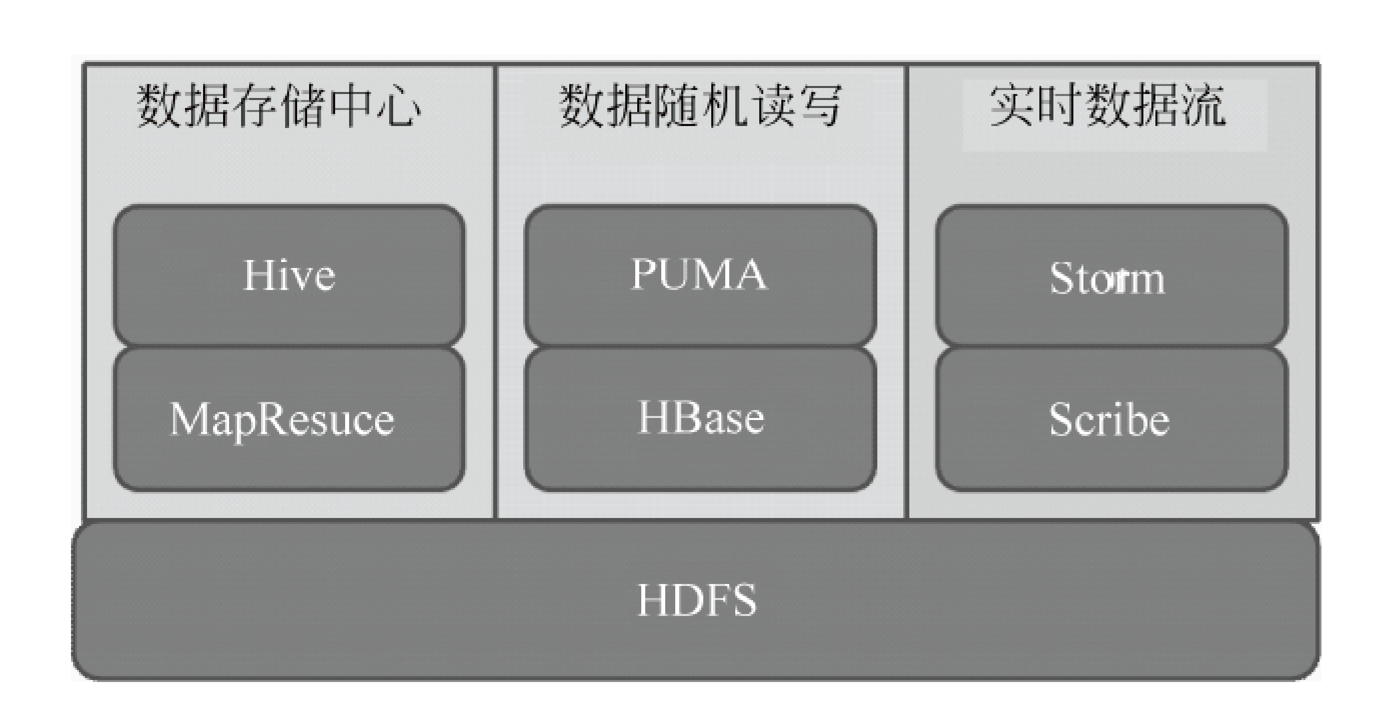


图 10-1　Facebook 的数据仓库架构[②]

再以电信运营商为例，说明 Hadoop 的大数据处理技术的应用。随着中国 3G 业务的普及和发展，中国三大电信运营商均将流量运营看作重要的收入来源，并大力拓展其移动网络市场。但流量增加的同时，3G 流量收费问题所带来的争议也越来越多。由于目前移动互联网主要以流量方式计费，且国内流量费用相对较高。一些

① 陈嘉恒，Hadoop 实战。机械工业出版社，2011 年 9 月。
② 董思颖，Facebook 开发的 HDFS 和 HBase 新特性，Hadoop 与大数据技术大会，2012 年 11 月。

不规范移动应用的自动更新等现象很可能会在用户不知情的前提下产生大量的费用，从而产生大量投诉。以中国联通为例，流量收费的投诉已占到总投诉的 7%～10%，并呈现迅速上升的趋势[①]。当用户投诉发生时，运营商需要向用户展示详细的流量使用情况，否则只能退费或赔偿。

但作为流量管道的运营商，其每天产生的用户访问日志远远超过任何一家互联网企业，且随着智能移动终端的普及仍在迅速增加。仅中国联通用户每半年的上网流量即可翻一番[②]。这种背景下，传统的流量计费设备根本已不可能支持海量日志的存储，更难以从大量日志中实时查询单个用户的流量使用状况。

目前，中国移动和中国联通均已在试点 Hadoop 解决方案。中国联通在北京、黑龙江、浙江和重庆四个地区尝试了 Hadoop+HBase 的解决方案，通过将用户的上网日志同步到 HDFS 中，并将用户上网的统计报表存储在 HBase 数据库中，通过 HBase 实现用户上网记录的快速查询。四个试点地区每月会产生 1200 亿条上网日志，只需 15 个数据节点的存储和计算资源，而从如此大规模数据中查询某个用户上网记录的响应时间不超过 1 秒。而这种问题，采用传统的解决方案基本上是无法解决的。

从上述例子可以看出，Hadoop 系统为企业面临数据规模急剧膨胀、对系统可靠性和实时性要求较高的应用提供了良好的解决方案。由于其良好的特性，目前已经在学术界和工业界受到广泛重视。多所科研院所对 Hadoop 集群展开了研究，其中包括斯坦福大学、CMU、加州大学伯克利分校等。国内的一些高校和科研院所，包括中科院计算所、清华大学、人民大学等也对其展开了研究，内容涵盖了存储结构、计算资源管理、任务调度、系统安全、HBase 等方面。Hadoop Summit 作为 Hadoop 社区的年度盛会，向人们展示了学术进展和商业案例。中科院计算所主办

① 大数据案例分析：电信业 Hadoop 应用分析，http://datacenter.watchstor.com/ news-138619.htm。

② 同上。

的 Hadoop 与大数据技术大会已成为中国 IT 界技术盛会，吸引了大批科研人员和知名企业参与。

在商业应用方面，Hadoop 技术已经在多个领域得到广泛应用，以满足企业存储和处理海量数据的需求。在使用者中，不乏像 IBM、Facebook、亚马逊、雅虎、推特这样的互联网巨头。国内包括百度、腾讯、淘宝、新浪、搜狐等在内的主要互联网公司均搭建了自己的 Hadoop 服务集群。除了互联网行业，在线旅游、能源开发、图形图像处理、医疗保健等多个领域的应用都逐步展开。美国国家航空航天局也在采用相关技术和系统处理包括星空图像在内的庞大数据。

那么 Hadoop 技术具体包括哪些内容？其作用分别是什么？图 10-2 画出了当前 Apache 框架下 Hadoop 相关项目的组成结构，不同的子项目适用于大数据处理中的不同场景，项目之间是互为补充的关系。

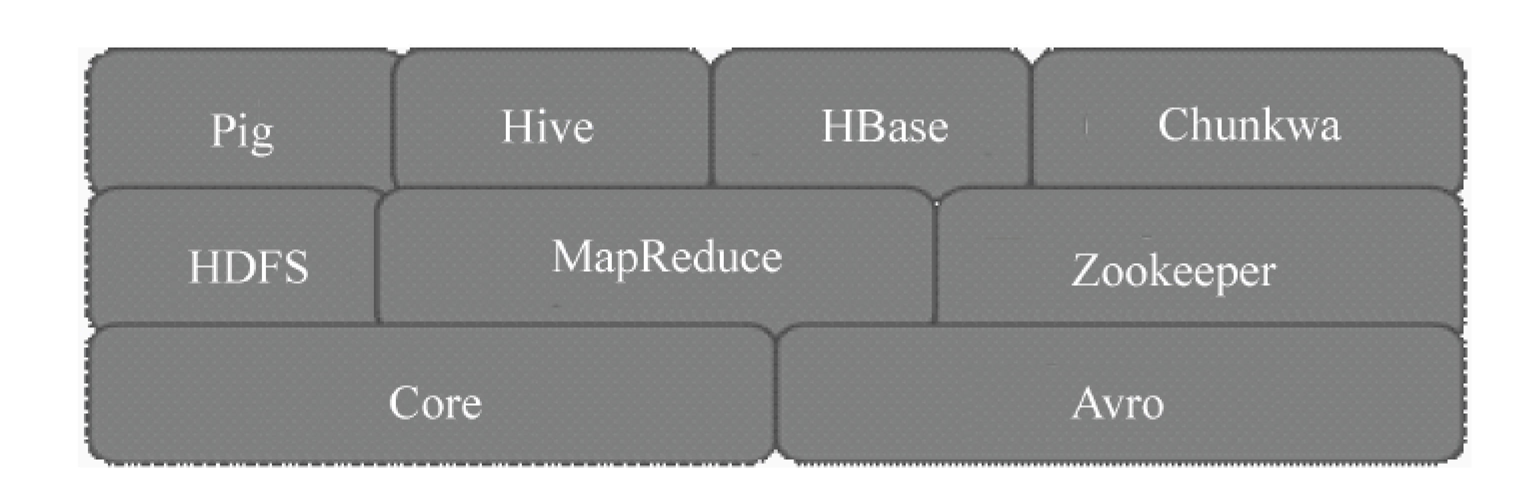


图 10-2 Hadoop 子项目组成[①]

1. Core 和 Avro 为底层支持框架，其中 Core 提供了一系列分布式文件系统和通用 I/O 的组件和接口，Avro 是一个高效、跨语言的数据序列号工具，用于数据的持久化存储。

2. HDFS 是一个块结构的分布式文件系统，用于集群中数据的存储和管理。

3. MapReduce 提供了一种并行处理大规模数据的编程逻辑。

4. Zookeeper 主要用于分布式数据管理的服务框架，如统一服务命名、集群管理等。

---

① Tom White 著，曾大聃、周傲英译，Hadoop 权威指南，清华大学出版社，2010。

5. Pig 是基于类 SQL 语言进行海量数据检索的数据流语言编程平台。

6. Hive 是基于 Hadoop 的数据仓库工具，可以将结构化数据映射为一张数据表，并提供 SQL 查询功能。

7. HBase 是一种基于 HDFS 的分布式、列式数据库系统，其实现原理是基于 Google 的 Bigtable①。

8. Chukwa 是一种基于 HDFS 的分布式数据收集和分析系统。

这些项目涵盖了基本的文件操作、基于分布式文件系统的分布式编程模式、数据库操作及数据分析等方面，形成了一个相对比较完善的大规模数据管理和处理体系，以满足不同的业务需求。

图 10-3 所示为 Hadoop 的体系结构图。在 Hadoop 中有三种不同的角色：客户端(Client)、名称节点（NameNode，也称元数据节点）和数据节点(DataNode)。

客户端可以通过应用程序对 Hadoop 中的文件进行创建、删除、移动等操作。除了具体操作命令不同，对客户端来说 HDFS 与传统的文件系统没有什么区别。

名称节点是 Hadoop 的核心节点，负责整个文件系统的管理和协调工作，其功能包括四个方面：① 元数据和文件块的管理—名称节点保存了文件基本属性、每个文件所对应文件及存储位置信息等。根据这些信息系统可以快速定位到客户端所需要处理的数据，是文件系统最重要的基础数据，因此称为元数据(Meta-data)；② 文件系统命名空间管理，记录文件系统元数据被修改的情况；③ 监听并响应客户端和数据节点的请求—客户端的任何操作，包括命名空间的创建与删除，文件的创建、删除和修改等，以及数据节点的文件块信息变化心跳响应等事件，均由名称节点统一调配、响应；④ 心跳检测—数据节点要定期向名称节点发送心跳信息以表明该节点仍然处于活跃状态，对长时间未检测到心跳信息的节点会认为出现故障并执行故障处理逻辑。作为 HDFS 核心的名称节点通常只有一个，如果发生故障将会出现系

① C. Fay, D. Jeffrey, G. Sanjay, H. Wilson C, W. Deborah A, B. Michael, C. Tushar, F. Andrew. "Bigtable: A Distributed Storage System for Structured Data". Research Google, 2006.

统瘫痪，甚至丢失所有数据文件。为保证名称节点的健壮性，有时会将其设计为主从结构，当提供服务的主节点失败时可以立即切换到从节点，从而不影响服务的正常运行。

数据节点是基本的数据存储单元，负责文件内容的存储。由于大集群中硬件故障是常态，为防止硬件故障带来的数据丢失，HDFS 采取了冗余复制的策略。文件会被分为不同的数据块，每个数据块被复制到多个数据节点中（默认为三份）。它将每个数据块存储在本地文件系统中，并保存文件块的元信息，同时周期性地将所有元信息发送给名称节点。

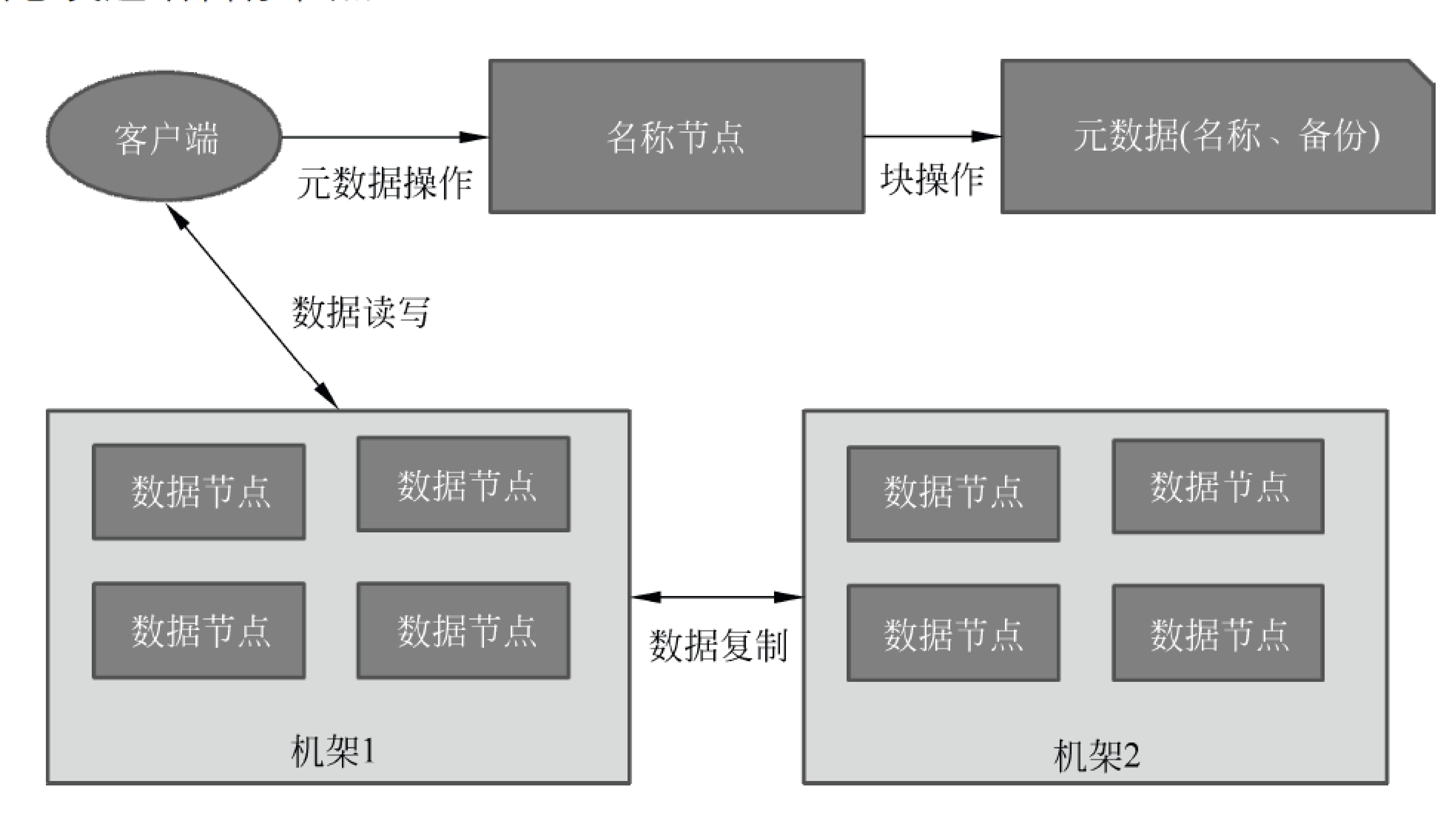


图 10-3　HDFS 体系结构图

第二节　数据挖掘技术和流程

Hadoop 系统的发展解决了企业大数据的存储和处理能力的问题。但是系统本身并不能对数据进行分析和理解。如何从海量的数据中发现有用的知识并为企业发展提供帮助和指导，是数据挖掘技术的研究目标。

简单来说，数据挖掘就是利用人工智能、机器学习、统计学、模式识别等技术，从大量的、含有噪声的实际数据中提取其中隐含的、事先不为人所知的有效信息的过程。一方面，数据挖掘所处理的数据对象是真实的、包含噪声，因此是一门实际应用科学；另一方面，其目的在于发现人们感兴趣的知识，与市场逻辑存在着紧密联系。大数据时代的数据挖掘技术并不是一门新的学科，其基本原理与传统数据挖掘并无本质区别。只是由于所需要处理的数据规模庞大、且价值密度低，在处理方法和逻辑上被赋予了新的含义。例如，传统数据挖掘由于数据量较小，为真实反映实际情况，需要构建相对复杂的模型；而大数据时代提供了海量的数据，即使使用相对简单的模型也可以满足需求。

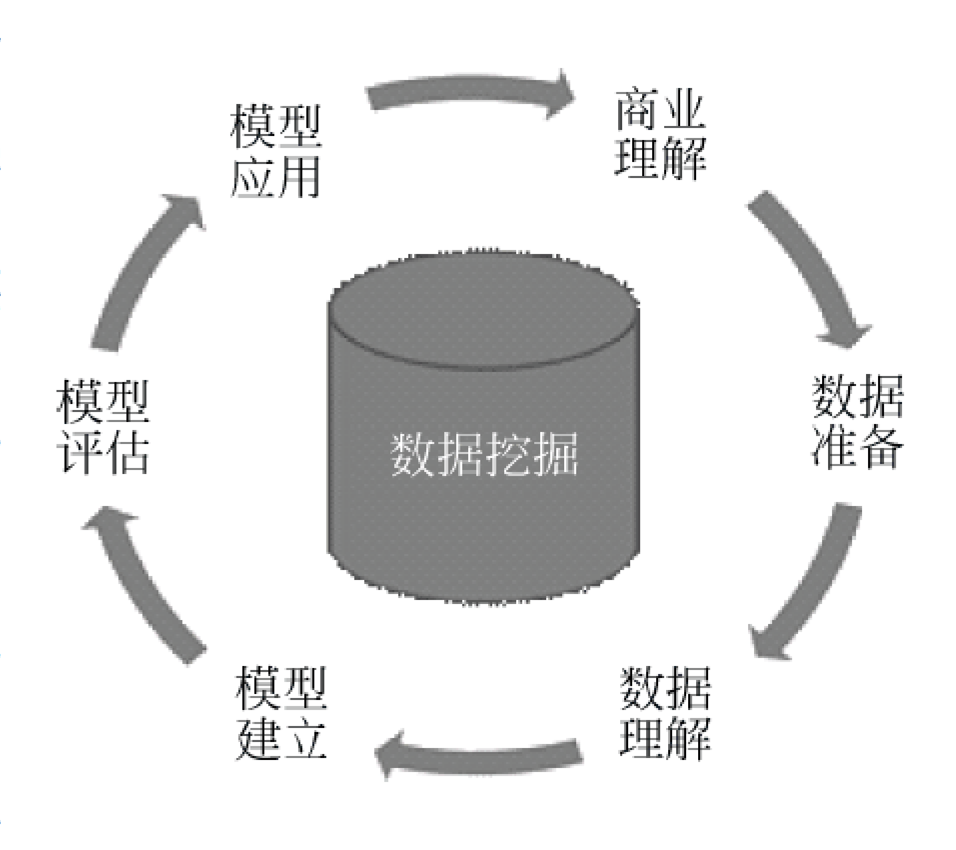


图 10-4　数据挖掘基本流程

图 10-4 所示为数据挖掘基本流程，包括商业理解、数据准备、数据理解、模型建立、模型评估和模型应用几个步骤。

首先是商业理解，也就是对数据挖掘问题本身的定义。所谓做正确的事比正确地做事更重要，在着手做数据模型之前一定要花时间去理解需求，弄清楚真正要解决的问题是什么，根据需求制定工作方案。这个过程需要比较多的沟通和市场调研，了解问题提出的商业逻辑。在沟通交流过程中，为了便于对沟通效果进行把控，可以采取思维导图等工具对沟通结果进行记录、整理。

明确需求后，接下来就是要收集并整理数据建模所需要的数据。这个过程是资源调配的过程，需要与企业的相关部门明确可以使用的数据维度有哪些，哪些维度与建模任务相关性比较高。这个过程通常需要一定的专业背景知识。

数据理解指的是对用于挖掘数据的预处理和统计分析过程，有时也称为 ETL 过程。其主要包括数据的抽取、清洗、转换和加载，是整个数据挖掘过程最耗时的过程，也是最为关键的一环。数据处理方法是否得当，对数据中所体现出来的业务特

点理解是否到位，将直接影响到后面模型的选择及模型的效果，甚至决定整个数据挖掘工作能否完成预定目标。该过程需要有一定的统计学理论和实际经验，并具备一定的项目经验。

模型建立是整个数据挖掘流程中最为关键的一步，需要在数据理解的基础上选择并实现相关的挖掘算法，并对算法进行反复调试、实验。通常模型建立和数据理解是相互影响的，经常需要经过反复地尝试、磨合，多次迭代后方可训练出真正有效的模型。

模型评估是在数据挖掘工作基本结束的时候对最终模型效果进行评测的过程。在挖掘算法初期需要制定好最终模型的评测方法、相关指标等，在这个过程中对这些评测指标进行量化，判断最终模型是否可以达到预期目标。通常模型的评估人员和模型的构建人员不是同一批人，这能保证模型评估的客观性、公正性。

最终，当挖掘得到的模型通过评测后可以安排上线，正式进入商业化流程中。为了避免由于建模数据与线上真实情况不一致而导致模型失效的状况出现，通常在应用过程中采取 A/B 测试的步骤，对模型在实际线上环境中的运行状况进行观察跟踪，确保模型在线上环境中符合预期。

了解了数据挖掘的基本流程，常用的数据挖掘任务和所用到的挖掘技术有哪些？总的来说，数据挖掘任务可以概括为描述性和预测性两大类。描述性任务主要是对现有数据进行理解和整理，从中发现其中的一般特性，是对历史知识的总结和归纳。预测性任务则是利用当前数据对事务的未来发展趋势进行推断，是知识的外延和推理过程。

比较常见的数据挖掘技术有以下几类：

1. 关联规则分析：包括频繁模式挖掘、序列模式挖掘，用于发现能够描述数据项之间关系的规则。典型应用是用户购物篮分析，发现用户经常一起购买的商品集合，如购买啤酒的人经常也会顺手购买小孩尿布；用户购买某商品之后后续最有可能购买的其他商品，如用户购买自行车两个月左右后通常会再购买打气筒。前者可

以用来指导商场的商品陈列，将用户最可能在一起购买的商品摆列在一起；后者则可以用来对用户的未来消费行为进行推荐引导。

2. 分类和预测：分类是按照已知的分类模式找出数据对象的共同特点，并将样本划分到相应的类别中，是最为基本的数据挖掘技术，广泛用于客户喜好分析、满意度分析等场景。如银行根据用户的消费能力和还款记录对其信用评级进行划分等。预测是将样本映射到连续的数值型目标值，从而发现属性间的依赖关系。例如，对产品未来一段时间的销售状况进行预测等。

3. 聚类分析：将一组对象按照相似性和差异程度划分到几个类别，使同一类别中样本的相似性尽可能大。例如，在金融行业中对不同股票的发展趋势进行归类，找出股价波动趋势相近的股票集合。

4. 推荐技术：根据用户的兴趣特点和历史的行为，向用户推荐其感兴趣的信息或商品。其最为成功的应用是在电子商务网站中，向用户推荐其可能购买的商品，从而增加商品的销售规模并提高用户粘性。

5. 链接分析：根据样本或数据对象之间的关联，可以构建对象之间的链接网络。链接分析是指利用图论模型对这些链接网络进行分析挖掘的一系列技术，其中最为知名的当属谷歌通过分析网页之间的跳转关系对页面权威度进行排序的 PageRank 算法。

其他相关挖掘技术还包括孤立点分析、数据演变分析等。

上述挖掘技术均在互联网、金融、生物医学、零售业等多个行业和领域得到广泛应用，并为相关企业带来丰厚的收益。以下将通过具体行业案例说明数据挖掘技术的使用方法及其价值。

啤酒与尿布——沃尔玛的营销神话

“啤酒与尿布”的故事已经成了营销界的神话，人们对数据挖掘技术的了解也几乎都是从这个故事起步。世界著名零售连锁超市沃尔玛拥有世界上最大的数据仓库

系统，其中收集了各个连锁店一年多详细的原始交易数据。为了能够准确把握消费者的购物习惯，沃尔玛利用数据挖掘工具对其顾客的购物行为进行了购物篮分析。系统通过不同物品之间的关联分析，了解哪些商品被顾客一起购买。一个令人惊奇的发现是，“啤酒”与“尿布”这两个看似完全不相干的商品经常出现在同一购物篮中。这一结论是对历史数据统计挖掘的结果，体现的是数据的真实情况。实际情况是这样的么？这一发现会给沃尔玛带来什么样的有用价值？

通过市场分析人员的调查发现，这一现象发生在年轻父亲身上。原来，在美国很多有婴儿的家庭中，通常是母亲在家照顾婴儿，年轻的父亲下班后去超市买婴儿尿布。父亲在买尿布的同时，也会顺手为自己买些啤酒。同时，如果年轻的父亲只能在卖场买到二者之一，他很可能会放弃购物而转去另一家超市，直到可以同时买到两种商品。

通过对海量购物行为的挖掘分析，沃尔玛不仅发现了一种有趣的现象，还偶然揭示了隐藏在其背后的每个人的一种生活模式。面对这一奇特现象，沃尔玛该如何应对？通过将啤酒与尿布并排摆放在一起使二者的销量双双增长。

按照常规思维，很难想象啤酒和尿布这两种商品会存在任何的逻辑关联。若非借助数据挖掘技术对海量购物行为数据进行分析，沃尔玛几乎很难发现数据中的这一内在价值，而这一关联关系的发现离不开关联规则分析技术的发展。常用的关联规则分析算法为 Apriori 算法、FP-tree 算法等。其中 Apriori 算法的基本原理构成了后续其他所有关联分析算法的理论基础，其基本挖掘流程如 10-5 所示。

这一故事带来的启示是什么？首先，为什么沃尔玛能够发现二者之间的关系？一个很重要的原因在于沃尔玛对数据技术的重视。其构建了先进的数据仓库系统对上千家卖场产生的海量用户购物行为进行分析，该系统不仅提供了庞大的数据存储能力，还需要具备强力的数据运算和挖掘能力。其次，该案例充分展示了数据挖掘技术在帮助把握用户购物习惯、帮助改善用户购物体验从而提升企业营销能力中所具备的巨大潜力。



图 10-5　利用 Apriori 算法进行关联规则挖掘流程图

啤酒与尿布的故事开启了数据挖掘技术在零售业应用的先河。目前越来越多的卖场开始重视对用户购物篮的分析，并用于指导其商品陈列、消费引导等。当然，不同类型的商场需要从购物篮中挖掘的内容可能有所区别，需要视企业需求而定。例如，日本的 7-11 便利店通常面积很小，所有的商品都陈列在相对狭小的空间中，简单的商品关联分析可能对其价值不大。但如果能够通过购物篮发现气温对碳酸饮料、凉面等的销量影响，或盒饭购买客户群的特点、购买时间分布等信息，无疑会更有价值。从这个角度，也说明了商业理解在数据挖掘流程中的重要作用。

计算广告系统——永不停息的印钞机

尽管服务种类五花八门、涵盖了生活娱乐商业等各个层面，互联网的盈利模式总结起来就是广告、销售、渠道三种。其中销售主要指的是电子商务，通过开展 B2B

或 B2C 甚至 C2C 业务帮助满足终端用户的消费需求；渠道主要通过向终端用户提供游戏、会员服务、信息增长服务等向目标用户收取相应的服务费。而应用最为广泛、最为直接的当属广告模式。包括谷歌、Facebook、雅虎、百度、淘宝等在内的主要互联网公司均依靠广告服务作为其主要收入来源。2012 年中国网络广告市场规模达到 750 亿元，并仍以每年超过 40%的速度迅速增长①。

互联网广告业务经过十多年的发展，在计费模式、广告形式、投放技术等方面都取得了巨大发展。当前广告投放主要包括以谷歌搜索为代表的搜索广告和以雅虎为代表的展示广告两类。由于用户的搜索词表达了当前强烈的查询意图，能够比较准确的把握用户兴趣点，因此搜索广告在过去取得了更快发展。但随着精准定向技术的进步，展示广告也取得了长足发展。根据用户访问页面的内容及用户的兴趣特点进行广告定向投放，可以使投放的广告更加契合用户的关注点，从而有助于提高广告点击率，增加广告投放效果，有助于整个广告产业链的良性发展。而实现广告精准定向投放的技术称为计算广告学（Computational Advertisement），其核心内容在于对页面内容的分析，用户的兴趣、性别、年龄等属性信息的挖掘，广告检索以及广告点击率、转换率的预测等，所使用的技术则涵盖了数据挖掘领域的绝大部分研究内容。

正是依赖于计算广告技术，互联网广告可以根据用户特点进行个性化广告展示，真正实现“千人千面”的营销目的，从而显著提高企业营销的针对性，使其成为区别于传统广告行业的新兴市场。同时，随着相关技术的日益成熟，根据网络流量的质量及其与广告主的相关度进行实时流量买卖成为可能，互联网广告也先后出现了网络联盟、广告交易平台、需求方平台、销售方平台及数据管理平台等角色，广告投放效果不断优化，形成了一个深度细分的产业链。

图 10-6 所示为计算广告相关的研究内容，涵盖了用户行为分析、广告检索、

① http://www.cnad.com/html/Article/2013/0106/20130106175119592.shtml。

广告点击率预测、拍卖理论、广告计费及欺诈检测等多个方面，且研究内容仍在不断丰富和深入，形成一个庞大的技术体系。

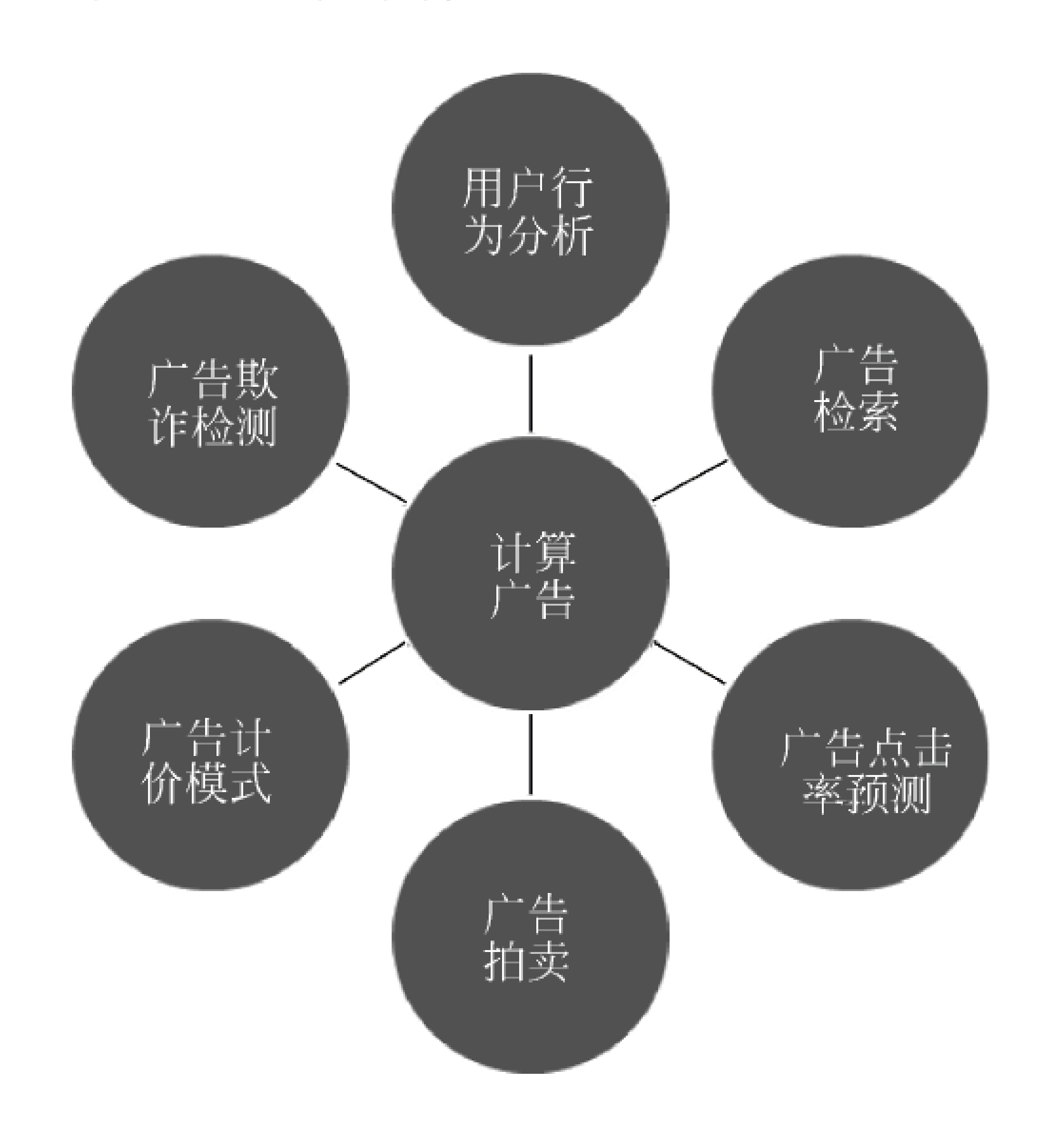


图 10-6 计算广告学的技术体系

以用户行为分析为例。通过分析用户浏览过的页面、所使用过的搜索词及其他的社交、分享、收藏、购买等行为，对用户进行分类和建模，把握用户的特点、兴趣及访问意图等，然后有针对性地投放广告。如果系统能够准确了解到用户是一个年轻妈妈，则向其投放婴儿用品或育儿教育相关的广告显然更能符合用户身份和广告主的营销目的。相反，向一个未婚男青年投放女性时装广告则不会取得好的效果。

在计算广告相关的最新进展中，最为引人注目的当属移动广告和广告交易平台两种。随着移动智能终端的迅速普及，2012 年移动广告的市场规模已迅速扩大至 50 亿美元。谷歌、Facebook、苹果等均积极部署其移动广告战略。而随着实时竞价技术的发展，广告交易平台逐渐成为新兴的广告投放渠道，其将每次广告投放机会都拿到一个公开交易市场进行拍卖，帮助广告主选择合适的投放机会。完整的广告交易链涉及需求方平台、销售方平台、交易平台、数据管理平台等，其中需求方平台代表广告主利益，帮助广告主购买合适的投放机会，代表公司为 MediaMath、

InviteMedia 等；销售方平台代表广大媒体的利益，帮助媒体实现每次投放机会的公开售卖，代表公司为 Admeld、Right Media 等；交易平台帮助需求方平台和销售方平台实现买卖交易，类似今天的股票交易所，代表为谷歌的 DoubleClick、淘宝的 Tanx 等；而数据管理平台则是帮助交易各方管理数据，通过强大的数据分析挖掘技术从而提高对这些数据的理解，代表为 BlueKai、eXelate 等。

在大数据时代，借助以 Hadoop 为基础的数据管理工具及相关的数据挖掘技术，计算广告系统已取得了很大成功，也面临着巨大的技术挑战。首先，是所处理数据规模急剧扩大，如谷歌的广告平台每天需要处理上百亿的广告请求，帮助成千上万个广告主实现广告投递，对系统的实时性和准确性都提出了越来越高的挑战。这也是大数据时代数据挖掘相关计算所面临的普遍问题。另一方面，计算广告也为互联网行业创造了巨大的市场价值，诸如需求方平台、数据管理平台等新的角色不断出现，市场分工越来越精细，所涉及领域逐渐向移动终端渗透，基于位置的广告、社交网络广告等新的技术层出不穷，向市场不断释放着新的投资和创业机会。

常用数据挖掘工具

到目前为止，数据挖掘技术已经取得了长足发展，许多数据挖掘商业或开源软件工具也逐渐问世，大大降低了数据挖掘的技术门槛。熟悉和掌握一些常用工具对日常的挖掘工作无疑能起到事半功倍的作用。此处简单介绍几种常见的挖掘工具，供读者选择使用。由于只是简单介绍，具体软件的使用，读者可以查阅具体的操作手册。

1. Excel：严格来说，Excel 并不是一款数据挖掘软件，但也集成了丰富的数据分析、数据挖掘、预测分析等方面的功能。同时，由于其广泛的应用范围、便捷的操作性和强大的数据处理能力，使其成为首选的数据挖掘工具。当数据规模不是很大时，可以使用 Excel 完成一些基本的关联分析、回归预测等任务。其缺点是能够处理的数据规模相对较小，且灵活性不足。

2．SPSS：最早的统计分析软件之一，提供了数据管理、统计分析、预测分析和决策支持等功能，其统计建模功能集成了方差分析模型、Logistic 回归模型等相对复杂的数据模型。其突出特点是操作界面友好，输出结果美观，能够以 Windows 视窗方式展示各种管理和分析数据，比较适合非专业人士使用。

3．Weka：一款新西兰怀卡托大学研发的开源机器学习和数据挖掘软件，在学术界得到广泛应用。几乎支持 Linux、Windows、Macintosh 等所有的操作系统，为普通用户提供了图形化操作界面，高级用户还可以直接通过 Java 编程对其进行扩展。Weka 中集成了非常全面的数据挖掘算法，涵盖了数据预处理、分类、回归、聚类、关联分析等多种模型。其缺点是对统计分析的支持相对较弱。

4．R：用于统计分析和图形化算法的编程语言和分析工具。与 Weka 类似，其源代码也开源自由下载使用，并支持多种操作系统。R 支持包括统计检验、预测建模及数据可视化等一系列分析技术。

5．Mahout：Apache 软件基金会开发的开源项目，是目前少数能够运行在 Hadoop 平台上的数据挖掘工具。已经实现了包括协同过滤、关联分析、分类、主题模型等在内的多种技术，但由于开发时间相对较短，目前每个领域所实现的算法相对较少。由于基于 Hadoop 平台实现，能够支持较大规模的数据处理能力。

其他常见的数据挖掘工具还包括 RepidMiner、Orange、LibSVM 等。

第三节　如何成为数据专家

Hadoop 技术的广泛应用及数据挖掘、数据分析技术的发展为企业低成本处理大数据提供了可能。但大数据技术的战略意义不在于掌握或拥有庞大的数据信息，而在于对这些数据的深度分析挖掘和专业处理能力，而这种能力的发育离不开数据人才的培养。市场环境越来越复杂，直觉式的市场决策已不能适应企业竞争的需求，越来越多的业务依赖于对内外部数据的深度解读。数据专家在未来企业竞争和战略

发展中将发挥越来越大的作用。因此，企业对数据专家的需求缺口越来越大。如图10-7所示为LinkedIn网站上对数据分析人才的需求趋势图。从图中可以发现，2000年之后企业对数据分析人才的需求量呈指数增长。

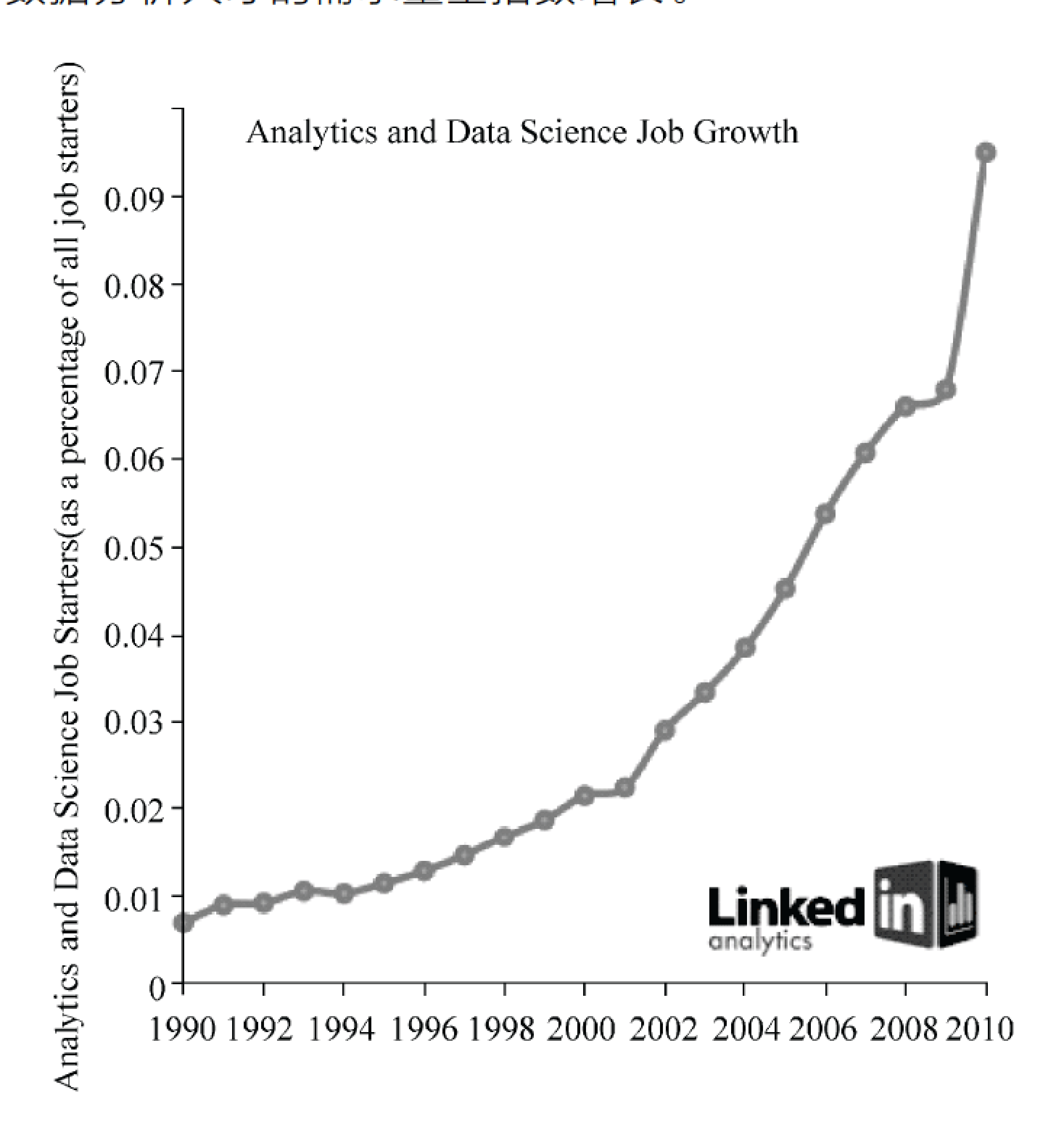


图 10-7　LinkedIn 数据分析人才岗位的增长趋势[①]

那么大数据时代，企业需要具备什么样的数据分析和挖掘能力？总的来说，企业对数据人才的需求涵盖了产品和市场分析、安全和风险分析及商业智能三个领域[②]。其中产品和市场分析主要用于了解行业发展状况、目标人群特点及新产品的市场接受程度等；安全和风险分析主要通过收集特定数据并进行分析，以发现并遏制网络入侵；商业智能主要对企业所拥有的杂乱无章的数据进行分析整理，从中挖掘有效

① Cashcow.企业需要什么样的数据科学家。http://www.ctocio.com/ management/career/5394.html。

② Cashcow.企业需要什么样的数据科学家。http://www.ctocio.com/management/career/5394.html。

知识，帮助企业进行业务经营决策。

从上述企业需求来看，大数据时代的数据专家与传统意义的分析师相比并无特殊之处。只是由于对数据的解读和把握能力与企业的业务联系更加紧密，对数据分析的及时性、准确性和深度等提出了越来越高的要求。同时，由于数据规模庞大和价值密度低等特点，对数据专家的个人素质和能力也提出了更多挑战。好的数据专家能够熟练使用复杂有效的统计技术和友好的工具软件对海量数据进行深度处理。而为了具备这种技能，企业和数据专家本身都需要持续的培训和学习。

可以从技术技能和个人素养两个角度来定义一个优秀的数据专家需要具备的能力。

技术技能

1. 扎实的统计学基础：统计学是当前很多数据分析和数据挖掘算法的理论基础。对统计理论，如概率分布、假设检验、贝叶斯理论等的理解有助于对数据进行更好的解读。

2. 深刻理解预测模型：能够使用常见的预测模型，如回归、聚类、决策树等在历史数据基础上对未来进行预测。对这些预测模型使用方法、应用场景的理解是数据专家的必备技能。

3. 熟练使用统计工具：为提高工作效率，数据专家需要熟练使用一种或多种分析工具。Excel 是当前最为流行的小规模数据处理工具，SAS 等工具也获得了广泛应用。而前面所介绍的以 Hadoop 为代表的数据管理工具将越来越广泛地应用于数据业务中。

内在素养

1. 数据敏感性：数据专家需要有很强的数据敏感性，能够处理大量的数字、Excel 表格甚至大型数据库。数据敏感性往往来源于类似问题的解决方法和经验，

并以这些经验为基础做出适当的调整。

2. 创造力：数据专家所面对的绝非一成不变的统计公式。每个业务所遇到的问题都是不同的，而解决这些问题的数据往往复杂多样，经常会遇到一些不可预见的问题。优秀的数据专家需要能够独立思考、富有创造力，能够根据具体业务需求寻找合适的解决方法。有创造力的人往往能够以很动听的故事描述其问题。

3. 强烈的求知欲：数据专家需要有从大批数据中挖掘有用信息的强烈的求知欲和好奇心，需要了解从何处入手，挖掘何种信息，提出正确的问题并坚持不懈查找答案。有时还需要对结果的深层原因进行穷追到底的深度挖掘。

4. 良好的沟通能力：数据专家经常会卷入各种商业行为中，需要与多个相关方共同工作。优秀的数据专家能够使用平白的语言对结果进行解释，并与相关人员打成一致。

一个初级的数据分析师可能只需要掌握基本的分析技巧便可胜任；成熟的数据分析师需要对数据分析方法有比较深入的理解；而资深的数据专家则应该具备丰富的经验和宽广的知识面，能够独立设计和完成相关解决方案。总之，随着企业对数据的重视程度越来越高，数据专家在企业经营和决策中所起的作用也越来越大，对数据专家的能力和个人素质均提出了更多要求。那么如何才能成长为合格的数据专家？根据上述对数据专家的能力要求模型，下面给出数据专家的成长路线图。

1. 端正价值观和职业操守：数据专家的职业特性使其有可能接触到企业的核心数据，有些甚至是只有企业少数高管才能看到。从企业信息安全和战略安全等角度均对数据专家特别是资深数据专家的职业操守提出了更高的要求。谦虚谨慎、务实自信的数据专家是企业不可多得的财富。

2. 了解生活常识，广其见闻：由于经常要面对海量的数据，并从中挖掘有价值的信息并排除异常数据，对数据专家的数据敏感性提出了很大挑战。所谓数据敏感性，通常是当看到数据后所能产生的直觉判断。而直觉的培养需要从积累大量的常识数据开始。

3. 建立合理的知识结构和深厚的知识底蕴：如前所述，数据的分析和挖掘工作是一门科学工作，需要具备深厚的统计学基础，并熟悉常用的统计模型。同时需要注意行业知识的积累。只有积累了足够多的知识底蕴，才能培养独到见解，为企业决策提供有价值的参考依据。

4. 经历大量实际数据项目的历练：实践出真知，任何理论的知识都要经过大量的实践才能转化为个人技能。资深的数据专家通常要经手多个数据项目，洞察各种数据内在的逻辑关联，具备纲举目张、一叶知秋的判断能力。

5. 熟练使用相关工具：数据分析和挖掘工作离不开相关工具的使用。资深数据专家需要对常用的数据工具熟练操作，从而从各个维度挖掘数据价值。

6. 重视团队价值：现代企业早已不是单打独斗的英雄主义时代，绝大部分业务的进展都渗透着团队合作的精神。身处数据业务核心的数据专家尤其要注重团队合作意识的培养，充分调动团队成员的资源和才智，重视对业务合作伙伴的支持。

第三部分

全景扫描

美国、欧盟、日本纷纷推出自己的大数据发展计划，并建立了 Data.gov 类的网站，促进数据的公开、分享。太平洋两岸活跃的新兴公司，则成为资本市场热捧的对象。政府、学术界、资本市场、产业界合力展开一幅万马奔腾、逐鹿中原的巨幕。他们正在重新定义世界！

导读：

1. 开放数据是政府建立“数字文明”的起点，促使政府更加高效、更加透明、更加廉洁、更加创新。大数据是一把双刃剑，同时也可能引起侵犯公民隐私、计划监控盛行的局面。简言之，大数据可以使政府更加开放透明，也可以加剧政府集权专制。如何选择取决于民族的智慧。
2. 透明和开放已经逐渐成为社会认可的价值观，但是具体到自己的领地，透明和开放就有可能变成一种威胁：以前的信息特权没有了，受到的监督和制约多了。因此要开放数据，就会有层层的审批和反复磋商，甚至故意的阻扰。这种文化几乎无处不在。美国也面临同样的问题，但在一系列法案的推动下，政府艰难但坚定地走上了透明和开放之路。
3. 奥巴马提出将致力于创建一个前所未有的开放政府的计划，推动政府成为公众能够信任的政府，成为公众能够积极参与、与之协作的开放系统。Data.gov 网站就是在开放政府计划的背景下诞生的。
4. 欧盟委员会全新的开放数据平台（以下简称为 ODP）Beta 版已经向公众开放（http://open-data.europa.eu/open-data），和美国政府的数据开放平台类似，致力于推动开放、透明的政府，促进创新。

第十一章

国家选择

大数据既能促使开放透明，又可加强集中管控。选择考验智慧！

——笔者

自 1993 年戈尔副总统提出“信息高速公路”计划以来，美国利用互联网技术，推动政府自身透明性方面的努力一直没有停息。2009 年 1 月 21 日，奥巴马就职后立即向联邦政府行政机构以及国家机构各部长发布《透明和开放的政府》备忘录。奥巴马总统的这个举动意义深远，2012 年发布的《大数据研究与发展计划》与其一脉相承。作为开放政府计划的一部分和具体执行单位，www.Data.gov 是全球第一个政府数据开放平台。美国政府利用该网络平台，公开政府的信息，鼓励政府和公众交流，以提高政府的效率；推动企业与政府合作，促进政府管理向开放、协同、合作迈进。只要不涉及个人隐私和国家安全的政府数据，均需由 Data.gov 全部公开发布。

美国开放政府计划和《大数据研究与发展计划》见附录三和附录四，本章主要详细介绍“www.Data.gov”——政府数据开放平台的价值。

Data.gov 把美国政府推向了一个前所未有的开放高度，提高了政府的效率，并聚集全社会的能量共同解决面临的复杂问题。创立之初，Data.gov 只有 47 组数据，截至 2012 年 12 月，已经有 37.9 万组包括地理数据在内的原始数据，横跨 180 个政府部门或下属机构。基于这些公开数据，政府已开发了 1264 个应用，社会开发了 236 个应用。其中，有关于健康生活的、有关于高效使用能源的、有关于教育的……在智能手机或电脑上，人们都可以自如地使用。

Data.gov 引导了全球政府开放数据的潮流。目前，已有 30 多个国家建立了政府数据开放平台，英国、日本、澳大利亚、印度等国都已经加入这一潮流，并享受它的好处。值得注意的是，在 Data.gov 众多海外访问者中，数量最多的来自中国。

第一节　Data.gov 的诞生

提要：

1.《信息自由法》是美国人民争取政府信息公开的重要成果之一。在此之后，一系列政府信息公开及获取的法律得以颁布，形成了

关于政府信息透明和公开的法律体系。

2. 开放政府是奥巴马利用网络力量推动政府形态转变的重要举措。它是以服务公众为导向，以网络技术为手段，转变过去条块分割、封闭的形象，塑造整体的、服务型的政府，致力于打造透明开放、高效参与、合作共赢的政府平台。

3. 提供高质量的数据是 Data.Gov 成功的基础。衡量高价值的标准包括以下任意一条：能够用于增强政府机构问责和反应的信息；能够提高公众关于政府机构及其运作的知识；该政府机构的核心长远任务；能够创造经济机会的信息。

从古至今，行政文化都有保密、封闭的基因。无论是在东方还是在西方，无论是历史上还是现在，政府首脑的第一反应往往都是安全为上，信息公开不如信息保密。现在，透明和开放已经逐渐成为社会认可的价值观，但是具体到自己的领地，透明和开放就有可能变成一种威胁：以前的信息特权没有了，受到的监督和制约多了。因此要开放一个数据，就会有层层的审批和反复磋商，甚至故意的阻扰。这种文化几乎无处不在。

美国也面临同样的问题，但在一系列法案的推动下，政府艰难但坚定地走上了透明和开放之路。

Data.gov 的法律和社会背景

《信息自由法》是规定美国联邦政府机构公开政府信息的法律，于 1966 年 7 月 4 日由美国总统林登·约翰逊签署，是美国当代行政法中有关公民了解权的一项重要法律制度。林登·约翰逊在签署时表示，这一法律保障人民在国家安全许可的范围内，能够获得一切（公务）信息，只能以国家的利益而不是官员个人的愿望判定何时需要限制情报公开。

1966 年以前，美国公民要想查阅政府部门的公务资料，经常被以“公众利益的需要”为由拒绝。当时人们可以依据美国 1789 年制定的《家政法》和 1946 年制

定的《行政程序法》的规定，向文献、档案的保存单位提出查询申请，但是在许多情况下，“公众利益”并没有具体的界定，政府官员经常滥用行政职权，动辄以“国家安全”、“政务机密”等理由，扣压本应公之于众或向申请人开放的资料和记录，任意扩大保密权限，官僚主义倾向迅速蔓延。

与此同时，美国的现代化进程在加快，经济、科技事务日益复杂，美国公众、民间社团以及经济界要求信息共享的呼声日渐高涨。联邦政府拥有丰富的信息资源，人们希望政府能够向公众提供更多、更好的信息服务。这种保密与公开的社会矛盾和反差，在 20 世纪 50 年代初期引发了关于“知情权”的调查、报道和宣传活动，这些活动为国会两院通过《信息自由法》奠定了坚实的舆论基础。

《信息自由法》规定了民众获得行政情报的权利和行政机关向民众提供行政情报的义务：联邦政府的记录和档案原则上向所有的人开放，但是有九类政府情报可免于公开；公民可向任何一级政府机构提出查阅、索取复印件的申请；政府机构则必须公布本部门的建制，本部门各级组织受理情报咨询、查找的程序、方法和项目，并提供信息分类索引；公民在查询情报的要求被拒绝后，可以向司法部门提起诉讼，并应得到法院的优先处理。这项法律还规定了行政、司法部门处理有关申请和诉讼的时效。

《信息自由法》是美国人民争取政府信息公开的重要成果之一。在此之后，一系列政府信息公开及获取的法律（见表 11-1）得以颁布，形成了关于政府信息透明和公开的法律体系。

表 11-1　美国联邦政府颁布的信息公开及获取的相关法律

名　　称	颁布时间	主 要 内 容
信息自由法	1966	行政信息公开为原则，不公开为例外；一切人都享有获得行政信息的权利
咨询委员会法	1972	公众有权查阅会议的记录、报告、草案研究或其他文件，在缴纳一定费用后，可以复制文件
隐私法	1974	解决政府信息公开与保护私人秘密两种制度的矛盾问题
阳光下的政府法	1976	规定合议制机关的会议必须公开，公众可以旁听会议，获得会议的信息

续表

名称	颁布时间	主 要 内 容
美国联邦信息资源管理政策	1985	明确规定了联邦机构收集、处理和传播信息，以及管理联邦信息系统与技术的总体政策指导方针
电子信息自由法	1996	作为信息公开的对象包括电子记录，规定了有效的公开措施等

正是在这套体系的保障下，美国公众才可以相对自由地获取政府信息。美国多年的经验也表明，美国政府信息公开、共享与开发，已经产生了很多社会和经济效益，促进了社会和私人领域创新，提升了社会福利。

奥巴马的开放政府计划

奥巴马上台时，提出将致力于创建一个前所未有的开放政府，推动政府成为公众能够信任的政府，并且是能够积极参与、与之协作的开放系统。Data.gov 就是在奥巴马开放政府计划的背景下诞生的。

开放政府是奥巴马提高政府运作效率的一种手段，是构建高效、廉洁、务实政府的基础，这至少具有以下三个方面的含义。

1. 构建透明的政府，即公众对政府事务有知情权，政府要及时告知公众政府在做什么。

2. 构建公众参与的政府，鼓励公众参与政府决策，监督或者协助政府提高决策和办事效率。

3. 构建协作的政府，创造各级政府、各部门、非盈利性组织、企业以及私人之间的互动条件，强调相互之间的合作。

开放政府是奥巴马利用网络力量推动政府形态转变的重要举措，是以服务公众为导向，以网络技术为手段，转变过去条块分割、封闭的政府形象，塑造整体的、服务型的政府，致力于打造透明开放、高效参与、合作共赢的政府平台。

Data.gov 是美国开放政府计划的旗舰级项目之一，其目的就是方便公众寻找、下载和使用政府部门高价值的机读数据。它不仅向公众提供了便利的信息获取途径，

更重要的是提供了一个公众能够分享信息的框架。

Data.gov 诞生并获得广泛的关注

尽管奥巴马已经表态要开放联邦政府的数据，让数据走出政府，得到更多的创新运用，但是联邦政府的各个部门还是非常忧虑，他们表达了各式各样的反对意见，害怕公众误读、威胁国家安全、冲突数据导致不信任……这些争论一时甚嚣尘上。

对一件复杂的事情，既然已经认定正确的方向，没有付诸行动而在一些问题上争论不清是毫无意义的。Data.gov 项目负责人昆德拉的心里很清楚，如果任由讨论继续下去，想要达成共识，获得实质性的结果基本无望，将导致项目流产。他相信只要坚持高价值的数据标准，从一些没有争议的数据开始，尽快推出一个分享平台，Data.gov 的成功就指日可待了。

提供高质量的数据是 Data.gov 成功的基础。衡量高价值的标准包括以下任意一条：

1. 能够用于增强政府机构问责和反应的信息。
2. 能够提高公众关于政府机构及其运作的知识。
3. 该政府机构的核心长远任务。
4. 能够创造经济机会的信息。

2009 年 5 月 21 日，距离奥巴马签署《透明和开放的政府》整整 120 天，Data.gov 上线发布了。Data.gov 初次上线只开放了 47 组数据，8 月 26 日，一次性新增了 178 项原始数据。

昆德拉不断完善 Data.gov 平台的功能，先后加入了数据的分级评定、高级搜索、用户交流以及社交网站互动等新的功能。例如，用户可以在网站上直接向联邦政府建议开放新的数据，而相关部门必须给出回应，若不同意开放，也要列出理由。

三年来，Data.gov 已获得广泛的社会关注，这些关注者来自世界各地。他们安装应用，浏览、下载和分析各类数据。Data.gov 正在帮助美国政府走在透明、开放和协作的路上。

如今，Data.gov 网站每月的访问客户数已经达到 15 万左右，并且在持续上升，如图 11-1 所示。

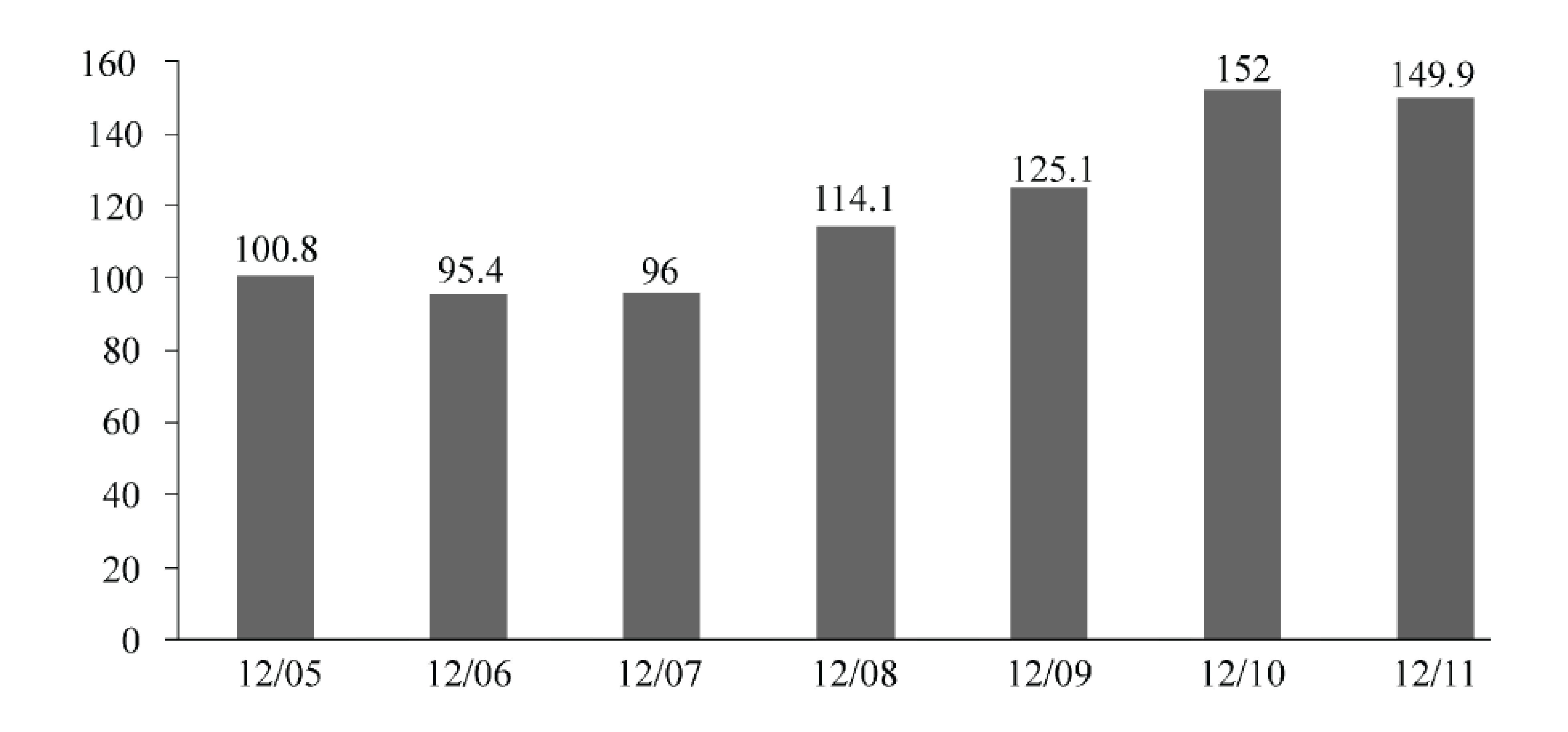


图 11-1　2012 年 5 月至 2012 年 11 月 Data.gov 访问客户数（万）①

每月的数据下载量已经超过了 5 万次，且在不断上升中，如图 11-2 所示。

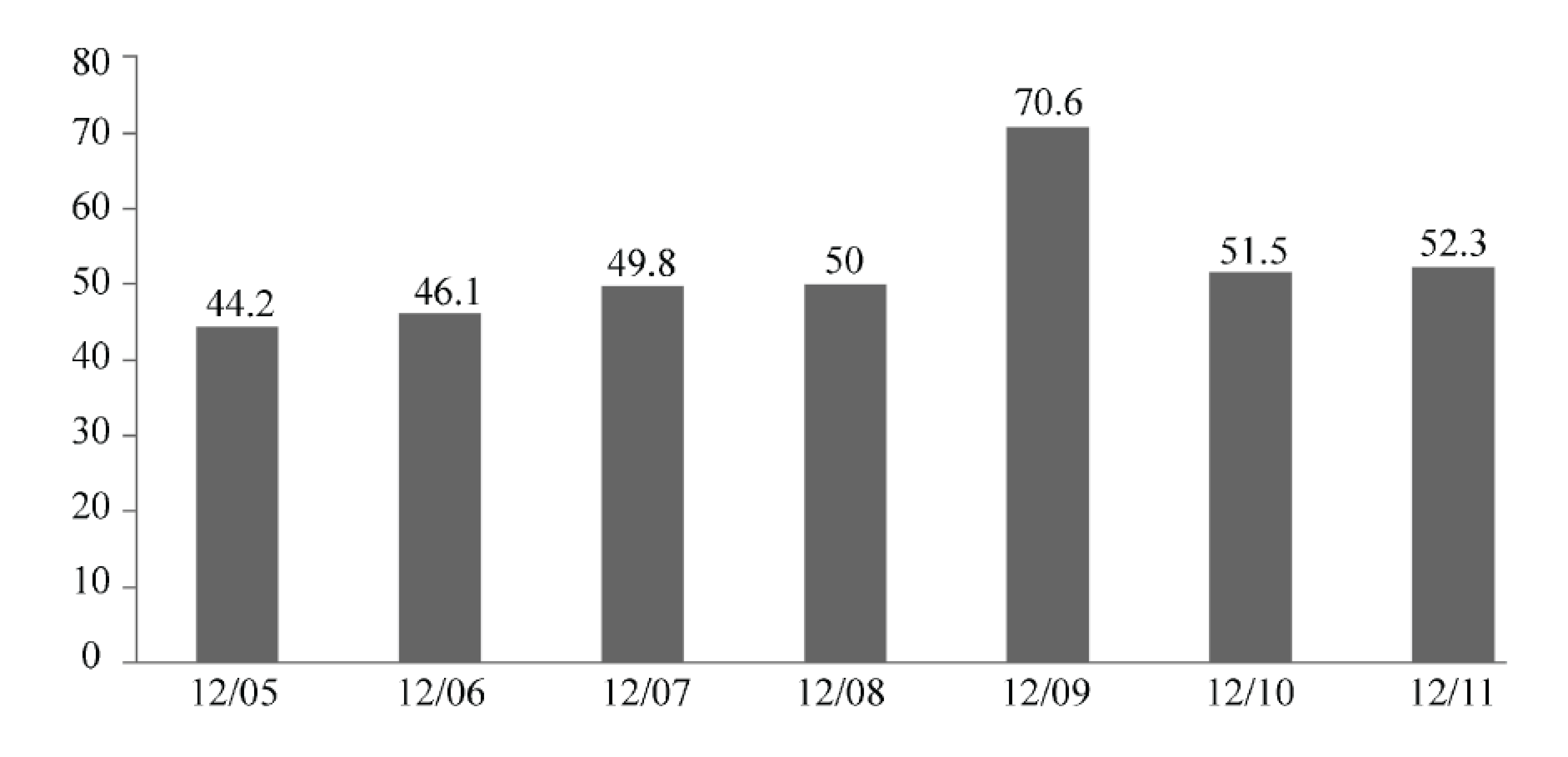


图 11-2　2012 年 5 月至 2012 年 11 月 Data.gov 数据下载次数（千）

公众下载的数据分类反映了当前社会关注的热点，见表 11-2。

表 11-2　2012 年 11 月份数据下载分类排序

数据类别	排序
地理环境	1
资讯和通信	2
劳动力、就业和工资	3

① 数据在 2012 年 5 月突然下降是因为统计方法改变了。本章引用的数据、表格如无特别说明，均来自 www.data.gov 网站。

续表

数据类别	排序
能源和公用事业	4
农业	5
国防和退伍军人事务	6
健康和营养	7
教育	8
执法、法院和监狱	9
金融和保险	10
联邦政府财政	11
出生、死亡、婚姻、离婚	12
商业企业	13
对外贸易和援助	14
国民收入、支出、贫困和财富	15
建设与住房	16
选举	17
自然资源	18
艺术、娱乐和旅游	19
国际统计数据	20
制造	21
经济	22

在 Data.gov 众多海外关注者（见图 11-3）中，最多的来自中国，反映了我国人民对信息的需求。

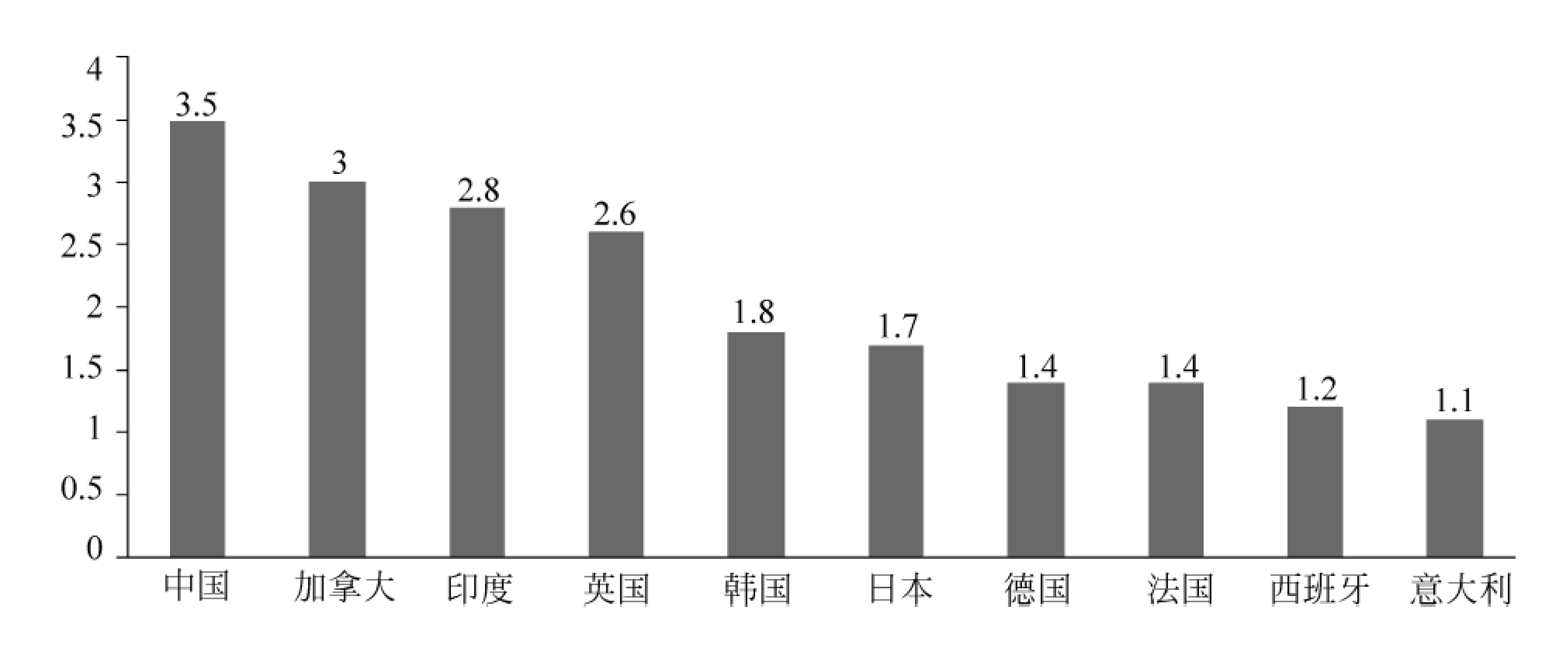


图 11-3　2012 年 11 月 Data.gov 前 10 访问客户来源国家（千）

第二节 Data.gov 的数据及应用

提要：

1. 公开原始数据是政府数据公开的关键，是公众能够再利用的基础。开放原始的数据集，社会公众就可以根据自己的需求去抽取、聚合、分析数据，挖掘有利的价值，使数据价值能够得到最大程度的发挥。
2. Data.gov 尽可能以原始数据的形式向公众免费开放，将分散的政府数据整合起来，减少了管理成本，有效地防范了欺诈与滥用，还创造了新的商机和就业机会。
3. 开源政府平台，带动全球政府开放数据。值得提示的是，访问 Data.gov 网站最多的海外关注者是中国人，足见国内科学研究、行业分析对数据需要的迫切程度。

Data.gov 网站发布了三类数据：“原始”数据集，用于快速查看、下载与操作系统无关的各种格式的机读数据；联邦数据集管理工具，提供各种数据摘录、抽取、分析工具，提供常用电子数据文件格式转换工具，以及标准的应用程序接口（API）；地理数据，美国政府信息中 80%的内容与地点有关，因而 Data.gov 提供综合地理数据，用户可以将这些数据叠加到地理基础信息上，生成地理空间信息服务。

“原始”数据集：不只是公开，还要让公众使用

是开放最原始的数据，还是经过加工和解释的数据？是粗线条的，还是粒度最小、最细的数据格式？

如果只是要求公开，政府就可以加工和解释数据，说明和解释的程度也在于政

府部门自己把握。但是这样数据的可用性就会大打折扣，往往被刻意地描述，只提供有利于政府部门的信息，甚至歪曲隐瞒某些关键的信息。在信息爆炸、数据泛滥的年代，美国人对于各种经过官方处理之后的统计数据都持有一定的怀疑态度。

因此，公开原始数据是政府数据公开的关键，是公众能够再利用的基础。开放原始的数据集，社会公众就可以根据自己的需求去抽取、聚合、分析数据，挖掘有利的价值，使数据价值能够得到最大程度的发挥。

同样，用最小的粒度把数据呈现给公众，让不同的用户各取所需。无论是警察还是居民，无论是企业还是慈善机构，都可以自己去决定怎样组合它们。可能的组合是无穷无尽的，这样数据才能发挥全部的潜在价值。

公众对政府数据的再处理，也是政府统计创造社会福利的过程。政府采集数据，如果只是经过加工后发布几个指数就束之高阁，完全是浪费数据价值。要想真正利用这一资源来实现对政府的责任监督，鼓励发明创新，那么从政府到企业到个人都必须有所作为。政府应该把数据公开变成一种常态，企业和个人也应该去推动政府公开原始和最小粒度的数据，并应用这些数据去创造价值。

Data.gov 尽可能以原始数据的形式向公众免费开放，将分散的政府数据整合起来，减少了管理成本，有效地防范了欺诈与滥用，还创造了新的商机和就业机会。

公众可以按照政府机构查找原始数据，其范围涵盖了广播理事会（Broadcasting Board of Governors）、商务部（Department of Commerce）、农业部（Department of Agriculture）、国防部（Department of Defense）等 180 个政府部门或下属机构的信息。

公众也可以按照数据分类查找原始数据，其范围涵盖了美国的人口、环保、教育、能源、地域、健康、法令等近 50 个种类的政府信息。

每一类数据都提供了详细的原始信息。比如失业统计数据，可以统计到国家、州、郡，可以选择具体的都市区、居住区、人口 25 000 以上的城镇等，数据内容

包括劳动力、在职人数、失业人数、失业率等（见表 11-3）。

表 11-3　佛罗里达州 2011 年 11 月至 2012 年 10 月的失业数据

年份	月份	劳动力	在职人数	失业人数	失业率
2011	11 月	9270820	8357474	913346	9.9
2011	12 月	9278251	8377933	900318	9.7
2012	1 月	9202927	8325213	877714	9.5
2012	2 月	9228929	8386739	842190	9.1
2012	3 月	9236237	8441031	795206	8.6
2012	4 月	9174350	8414434	759916	8.3
2012	5 月	9296361	8505322	791039	8.5
2012	6 月	9341306	8498250	843056	9
2012	7 月	9357791	8481237	876554	9.4
2012	8 月	9314152	8476972	837180	9
2012	9 月	9387109	8577256	809853	8.6
2012	10 月	9392740	8624024	768716	8.2

在数据格式方面，Data.gov 提供了多种机读数据下载格式，包括 XML、RSS、CSV/Text、KML/KMZ（地理空间数据格式）、ESRIShapefile（地理空间数据格式）等，保证了数据可用性。

当然即使政府有决心，也不能保证及时完整地公开数据。用户可以在网站上直接向联邦政府建议开放新的数据，2009 年 5 月至 12 月，Data.gov 共收到社会各界约 900 项开放数据的申请，其中 16%的数据立即开放，26%将在短期内开放，36%纳入开放计划，还有 22%因为国家安全、个人隐私以及技术方面的原因无法开放。

数据管理工具：让数据整合产生“1+1>2”的效果

Data.gov 提供各种数据摘录、抽取、分析的工具，提供常用电子数据文件格式转换工具。用户使用浏览器就可以快速地搜索、筛选各种数据，创建各种图、表，并自动转换成常用的电子数据文件格式，也可以通过 Widgets、Gadgets、RSS

feeds 等工具进一步自动获取数据。

Data.gov 还提供标准的应用程序接口（API），供开发者创建特色的应用。开放 API 是 Data.gov 战略的重中之重，每个交互式数据都有一个公开的接口，这些接口通过 http://explore.data.gov/catalog/next-gen 都可以查找到。Data.gov 同时还为开发者提供了教程和视频录像作为参考。

Data.gov 是一个政府数据的集散地，个人或企业可以通过交互式数据的 API 直接调用该网站的数据库，将不同类别的数据整合起来，创建新的应用，使“1+1>2”。

暴雨、风雪和其他天气相关数据，能够帮助商品零售企业预测需求，从而调整商品的储备；FDA（美国食品和药物管理局）的召回通知、相关药品的销售数据，也为诉讼律师、替代药品的供应商提供了潜在的机会。这些仅仅是把不同的数据放到一起，还没有做更多深度的聚合工作，就能建立新的见解，帮助公众创造效益。

美国历年的大豆生产面积、产量、价格、出口量、加工，以及豆产品销量、大豆作业机械销量、大豆主产区的天气、运输价格等数据来自不同的部门，将这些数据糅合起来，以行业分析框架为基础，就能建立起可靠的大豆行业分析模型，为豆农、豆油生产商、农产品期货等个人或企业创造可观的经济价值。

Data.gov 上有超过 30 万种可用的不同数据集，跨越了 50 多个种类，只要你能在它们之间建立联系，或者把它们与你的工作建立联系，就可以把它们聚合在一起为你工作。

仔细分析你的目标市场和客户，梳理出他们的特征，以及各种外部影响因素，积极的或者消极的，如气候条件、人口出生率、人口密度、老龄化情况、不同层次的零售商分布等，再搜索 Data.gov，找到相关的可用内容，你会发现网站提供了绝大多数的数据，即使没有的，也能找到可以替代的近似数据，这些就是你聚合原材料，创建应用直接调用它们的 API，如果你能准确地推断出相关的趋势或者行为，将会帮助你制定合理的商业策略。

对这些企业和个人来说，开放 API 简化了使用政府数据的环节，免去了维护数

据的成本，同时也降低了基于 Data.gov 的数据进行创新的门槛。

在商业领域，腾讯、淘宝、Facebook 等商业企业利用开放 API 完善了自身平台的功能，通过第三方的力量建立了自我完善的平台生态体系。同这些商业企业一样，Data.gov 将政府数据作为一种资源，开放 API 使整个社会都可随时调用、创新应用，挖掘政府数据的潜在价值，让 Data.gov 的平台价值自发壮大。

实施开放 API 战略后，Data.gov 发展思路也更加清晰了，只要做好关键的两项工作：保证原始数据的质量和数量及提供标准的 API。其他能让社会公众做的事情尽量让社会公众来做。

随着智能手机的普及，移动互联网的时代已经来临。Data.gov 也在推动开发相关的手机应用，把政府的公共服务搬到手机上。新战略颁布之后，奥巴马立即下令，每个联邦政府部门都必须在一年内推出至少两款利用手机提供公共服务的应用程序。

地理数据：提供直观的信息载体

以地图作为表达政府信息的载体，是当前国际社会普遍采用的方式。这是因为地图的表现更加直观，一张地图可蕴含上百万字的信息量。对于政府部门发布的信息，公众可能很难直接理解数据本身的含义，但通过具象化的地图，则能立即将数据与地理空间方面的经验结合起来，从而形成新的认知。

2012 年 12 月 14 日，美国康涅狄格小学校园枪击惨案震惊世界，许多美国人也希望美国政府加强枪支管制。美国纽约州《新闻报》在网络版上，发布了名为《隔壁的持枪者：你所不知道的街区武器》的文章，公开了韦斯特切斯特和罗克兰的“枪支许可地图”，读者点击后可获知每一名持有枪支许可居民的姓名和地址。这涉及到泄露私人信息的问题，但利用这个地图，也可以很直观地了解这些街区的枪支分布情况，帮助公众评估潜在的危险。

Data.gov 网站也已经以原始数据集为基础提供地图信息服务了。如果只提供数据，不免有被修改利用的可能，而如果提供的是基于数据生成的地图服务和有选择

性下载的数据，公众获取的就是真实有效的信息。Data.gov 一直在持续改进地理数据的体验，不断优化交互界面和底层结构。

点击“添加到地图（Add To Map）”，就可以通过地图，一目了然查看到关注的某方面信息。图 11-4 所示为美国活跃的飓风和热带风暴分布图，非常直观，利用这个信息，游客就可以很方便地规划行程，运输公司也可以规划线路避免不必要的损失。

地理数据也被用来创建各种有趣的应用。福布斯杂志网站利用 Data.gov 中人口流动数据（主要是纳税信息），开发了美国人口迁移的可视化工具。在该应用中，点击地图任意两个地点即可查看到这两个地点人口迁出和迁入的情况，企业利用这个信息作为决策参考，就可以做出准确地建立分支机构、营销资源投向等决定。

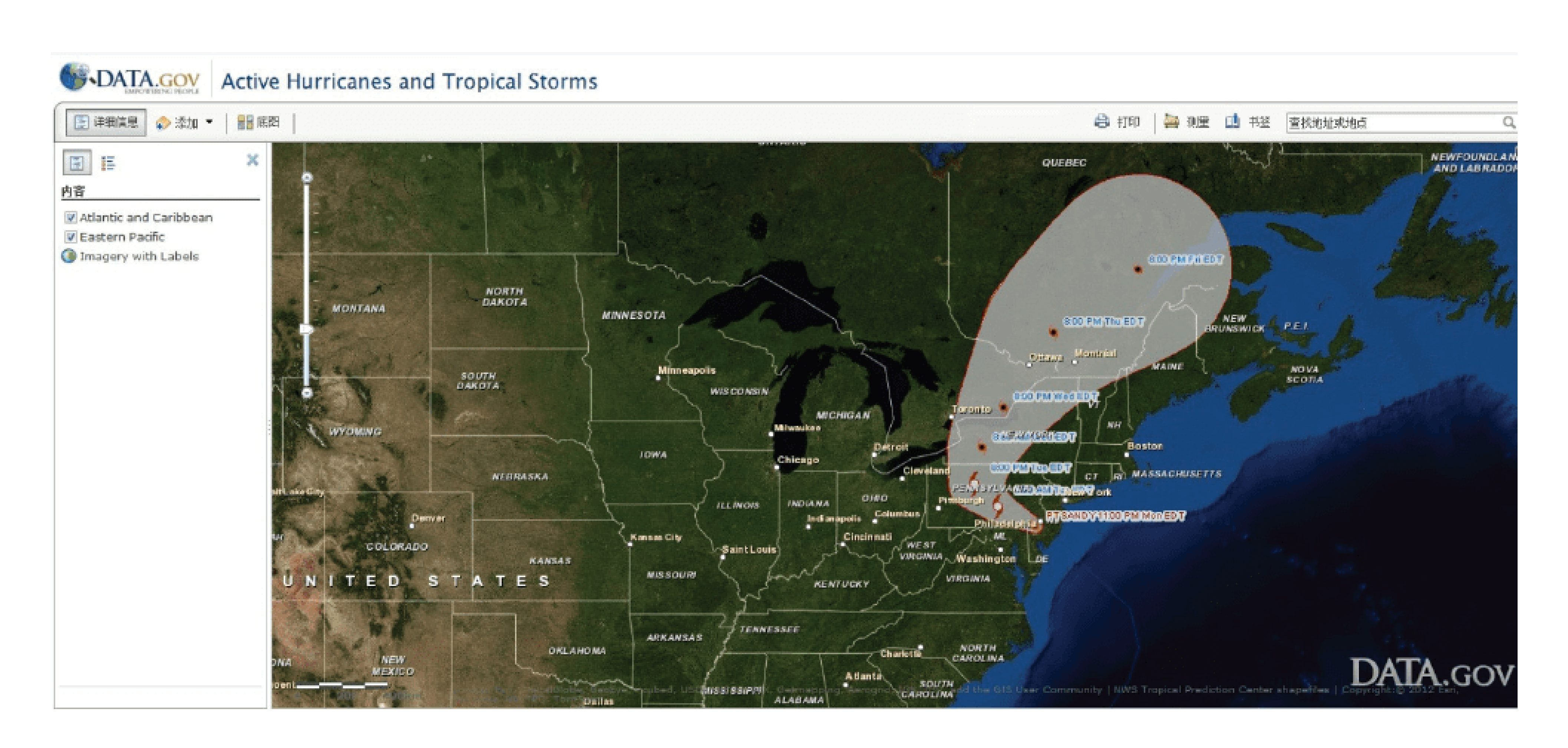


图 11-4　活跃飓风和热带风暴分布图

社区：创造分享信息和交流的机会

奥巴马的透明和开放政府计划，将 Data.gov 作为一个重要的激发创新的工具，因此提供了完善的互动功能，以创造分享信息、共同探讨问题的机会。

Data.gov 已经创建了商业、能源、教育、海洋、安全、供应链等 14 个专题社

区。那些对这些专题感兴趣的人，已经将 Data.gov 的社区作为重要的交流园地。来自学术界、产业界以及公众中的专家在这里讨论、探索问题，发布新的创意应用；政府也在这里公布需要解决的问题，利用 Data.gov 的数据，聚集社会力量共同创新，提供解决方案。Data.gov 的社区已在为美国社会创造各式各样的福利：激发社会创新、节约社会成本、提供企业家发现和利用新技术的平台、支持开发者马拉松（code-a-thon）比赛……

教育社区是教师、学生以及关心教育的应用开发者共同的家园。这里是各种教育资源聚集的中心，不仅有原始数据、可视化的工具，还有各种教学资料和应用，从摇篮时期的教育到职业教育都有丰富的材料。教师在这里可以找到课程素材、教学方案，还可以探讨激励学生的方法；学生在这里可以交流科学实验、课程论文，让自己的课程作业更加酷；应用开发者在这里可以挑战各类问题，开发教育类应用，提高教学或者学习的效率。

美国国家科学教师协会和美国能源部在教育社区邀请开发者开发“美国家庭能源教育”应用，并提供奖金支持。这个应用主要有三个目标：帮助美国 3～8 年级学生学习能源知识以及如何提高能源的利用率；帮助学生了解降低他们家庭能源消费的途径；激励学生和他们的家庭改变生活方式，减少能源浪费。

开源政府平台：带动全球政府开放数据

Data.gov 的“开源政府平台（Open Government Platform，OGPL）”由美国和印度政府合作开发。他们把代码托管到 GitHub[①]上，其他国家以及开发者可以直接使用代码，或者对代码进行修改以符合使用要求。

OGPL 主要包括以下四个核心模块：

1. OGPL 网站：政府机构用于发布数据、文件、服务、工具以及应用的模块。

① 网址：https://github.com/opengovplatform/opengovplatform。

2. 数据管理系统：用于向 OGPL 网站上载数据的模块。

3. 内容管理系统：用于管理和更新 OGPL 网站不同功能的模块。

4. 访客管理系统：用于与客户交互、反馈客户建议的模块。

Data.gov 引导了全球政府开放数据潮流。目前已有近 40 个国家、地区或者国际组织建立了开放数据平台，如图 11-5 所示。



图 11-5　建立了开放数据平台的国家、地区或者国际组织

第三节　开放数据是政府“数字文明”的起点

提要：

1. Data.gov 提升政府运作效率的效果非常明显，表现在三个方面：减少了大量政府信息系统重复建设；降低了公众获取政府信息的成本；有利于提高政府部门之间协作的效率。
2. 一些特殊部门或居于特殊职位的官员，掌握了国家的各种重要信息，具有信息特权，如果透明、公开，绝大多数的信息腐败就没有存在的可能，反而可以激发公众开发和利用信息资源创造价值。

3. Data.gov 的包容性打开了政府内各部门、政府与民众之间的边界，信息孤岛现象不再存在，数据共享成为现实。政府各机构开放创新：提供数据，提供问题和激励，邀请社会公众共同解决问题，通过众包的形式激发了大众的智慧，推动了社会创新。

Data.gov 带来政府数据资源共建共享的新理念，实现了政府信息由封闭和割裂向开放和整合的转变，同时也带来了政府信息公开由静态服务向动态服务转变，由单向服务向双向互动交流服务转变。

Data.gov 的透明、开放和协作机制，是高效、廉洁、创新政府的重要推动力，也是政府在日新月异、日益复杂的社会发展趋势中更有作为的前提。

高效的政府

创建高效的政府是奥巴马“开放政府计划”的一个目标，Data.gov 在其中起到以下重要的作用：

1. 减少了大量政府信息系统重复建设。美国总务管理局（General Service Administration，GSA）负责审批政府部门的软件开发项目，对那些有特殊需求的业务，如果开放政府平台的现有产品无法满足，GSA 才允许进行定制开发。这样确保了一次开发，多次受用，一家开放，多家受益，减少了不同政府部门的重复投资。开放政府平台作为“一站式政府云计算服务”，节约了大量 IT 预算。

2. 降低了公众获取政府信息的成本。企业和个人有获取“一站式”的政府信息服务的需求，原有的按职能划分的“多站式”服务让公众付出高昂的成本，非常不经济，也会导致政府和公众之间互相不理解。把各个政府职能部门的信息整合到一起，形成完整的政府信息服务产品，将有助于形成政府信息透明的氛围。Data.gov 提供 50 多类数据，以及处理这些数据所需的软件工具，所有人都可以自由下载使用，网站的数据资料不仅有利于公众了解政府政策，也对居民的日常生活起到实在的帮

助。一份由 16 万份行政区地图组成的精确到道路、建筑物、水系、行政区界线等详细资料的庞大美国地图，是网站上被下载最多的资料之一。

3. 有利于提高政府部门之间协作的效率。无论在哪里，政府只要有不同的部门，就会存在机构重叠、职责交叉、政出多门的矛盾以及权限冲突。政府部门间信息聚合、共享将有助于政府领导者或公众及时发现协作问题，帮助简化公务手续和环节，减少频繁的沟通和协同，提高行政效率。我国也在探索实行职能统一的大部门体制。实际上，健全部门间协调配合机制，可以从部门间信息的聚合、共享开始做起。

廉洁的政府

与金钱有关的腐败容易引起重视，但与信息有关的腐败因为更具隐蔽性，一般不容易被发觉，造成的危害也可能更大。政府部门掌握大量的信息资源，很容易被刻意利用，哪怕单独透露给需求该信息的企业或个人，也很难找到直接证据确定其违法违纪行为。但是，如果做到透明或公开，不仅可以抑制腐败，还可以创造更多的社会福利。

比如，城市规划信息中就有巨大的利益。如果政府没有及时公布新的区域或主干道路规划方案，被某个人员事先透露给有关联利益的房地产商，即使不投资，抢先买来地囤着，也能赚个盆丰钵满，受到损害的是政府和民众的利益。政府的财政预算、行业规划、政府采购、国资拍卖等信息，如果没有透明、公开的渠道，都有可能导致腐败。

当今世界正在向信息社会过渡，在信息社会中，信息成为比物质和能源更为重要的资源。政府掌握着大量的信息，一些特殊部门或居于特殊职位的官员，掌握了国家的各种重要信息，具有信息特权，如果透明、公开，绝大多数的信息腐败就没有存在的可能，反而可以激发公众开发和利用信息资源创造价值。

创新的政府

大数据给个人生活和企业环境带来翻天覆地的变化，大数据也将显著改变政府的作用和工作方式。

Data.gov 的包容性打开了政府内各部门、政府与民众之间的边界，信息孤岛现象不再存在，数据共享成为现实。政府各机构开放创新：提供数据，提供问题和激励，邀请社会公众共同解决问题，通过众包的形式激发了大众的智慧，推动了社会创新。

政府各部门也成为创新的主体。开展业务数据分析，发现数据背后隐藏的模式和微妙关系，揭示过去的规律，预测未来的趋势，创新工作方式，以制定更好的公共决策，用新思路、新方法、新举措破解经济社会发展过程中遇到的各种问题。

中国在将近 20 年的信息化建设中，投入了千万亿资金，沉淀了大量的宝贵数据，这些数据是整个社会经济活动的数字化记录，是可以无限次重复利用的特殊的非物质财富，是不可或缺的管理和决策的依据[①]。如何把这些数据聚合起来，提高行政效率和透明度，创新工作方式，提高对社会的服务能力，在这些方面，Data.gov 的道路值得借鉴。

第四节 欧盟开放数据平台——Open Data Portal

欧盟委员会全新的开放数据平台（以下简称 ODP）Beta 版已经向公众开放（http://open-data.europa.eu/open-data），和美国政府的数据开放平台类似，致力于推动开放、透明的政府，促进创新。

2010 年 4 月，欧盟委员会发起欧洲数字化议程，致力于利用数字技术刺激欧洲经济增长，帮助公众和企业最大化利用数字技术。ODP 是欧洲数字化议程的一部分，欧盟委员会副主席 Neelie Kroes 说："这将打开一个金矿，通过这个系统，公众获得这些数据会更便捷，成本更低，获得的数据内容更广泛"。

---

① 丁健，浅析大数据对政府 2.0 的推进作用，中国信息界，2012 年其 9 月。

截至 2013 年 1 月 12 日，ODP 已经开放了 5815 个数据集，其中的 5638 个数据集来自欧盟统计局 Eurostat，包括地理、大气、国际贸易、农业等各类信息。

ODP 提供的不仅是数据，还建立了数据的统一语法规则，保证数据发布机构、公众、应用开发者都能够利用这些数据，任何人都可以在这里下载数据，利用这些数据开发新的应用。

“原始”数据集

和美国政府的数据开放平台一样，ODP 开放最原始的、粒度最小的、未经过加工的数据，保证数据的真实性，让公众各取所需，各尽其用。

目前，ODP 数据提供 dft、sdmx 和 tsv 三种标准格式供下载使用。

表 11-4 是 2003 年至 2011 年关于欧盟成员国内陆货运水路的数据，其根据 2006 年 9 月 6 日欧洲议会通过的 1365/2006 号欧共体条例收集。这些数据是了解欧盟成员国货运情况的基础。

表 11-4　欧盟国家内陆货运水路　（单位：km）

国家	2003 年	2004 年	2005 年	2006 年	2007 年	2008 年	2009 年	2010 年	2011 年
保加利亚	4316	4259	4154	4146	4143	4144	4150	4098	4072
芬兰	5851	5741	5732	5905	5899	5919	5919	5919	5944
意大利	15965	15916	16225	16295	16335	16529	16686	16704	16726
立陶宛	1774	1782	1771	1771	1766	1765	1767.6	1768	1768
拉脱维亚	2270	2270	2270	2269	2265	2263	1884	1897	1865
荷兰	看作 1	2811	2810	2797	2801	2888	2896	3013	3013
波兰	19900	20250	20253	20176	20107	20196	20360	20228	20228
罗马尼亚	11077	11053	10948	10789	10777	10785	10784	10785	10777
斯洛伐克	3657	3660	3658	3658	3629	3623	3623	3622	3624
……	……	……	……	……	……	……	……	……	……

注：数据来源 HTTP://OPEN-DATA.EUROPA.EU/OPEN-DATA。

应用

对普通公众来说，原始数据集是难以阅读的。ODP 为帮助对数据感兴趣的公众浏览、理解和运用数据，提供了两个数据可视化应用，CubeViz 和 SemMap。

CubeViz 提供用户友好的界面。用户可以定制数据，选择不同的图表去展现数据，如饼图、直方图、曲线图、面积图、散点图……通过这些图表，更直观地去理解数据。

SemMap 是一款地图类应用。用户可以筛选感兴趣的与地理相关的数据，在地图上展示，将数据与对地理空间的理解结合，获得新的知识。这与美国政府的数据开放平台的地理数据类似。

任何人都可以基于 ODP 的数据和接口开发新的应用，也可以申请在 ODP 上发布自己的应用。

链接数据

ODP 的链接数据使用了标准的语义网技术，在这里可以查询“原始”数据集的元数据。

SparQL 就是 ODP 提供的查询终端，它基于 ODP 中数据组的元数据汇总表去查询。元数据汇总表采用高级数据管理系统的普遍规则，以保证可用性。

ODP 目前还是 Beta 版，相比于美国政府的数据开放平台，其提供的信息有限，功能也不够完善，但是从欧盟委员会的长远规划、对此寄予的厚望看，ODP 任重道远，它将是欧洲开放、透明、创新的重要催化剂。

导读：

1. 老牌企业如 IBM、甲骨文、EMC、微软，它们凭借在信息科技领域的积累和沉淀，依然具备举足轻重的作用。但是产业变化速度和发展趋势已经脱离这些巨头的控制，在更广阔、更宏大的经济社会舞台上狂奔。
2. 新兴巨头如谷歌、亚马逊、苹果、Facebook，它们之间的合纵连横，将决定未来十年的产业生态和竞争格局。它们拥有不同的商业模式、不同的盈利要素，但却都拥有同样规模庞大的数据资产。正是这些各具特色的数据资产，决定了它们的商业模式和未来走向。
3. 以这些庞大的数据资产为核心，新兴巨头有可能演变成为整个社会信息基础设施的一部分，具备颠覆其他产业的主导力量。而那些仅仅具备技术无法累积数据资产的公司，很可能成为它们的附庸，失去信息科技发展主导权。

第十二章

巨头碰撞

新兴巨头，正在重新定义产业生态和竞争格局，老牌公司沦为看客。

——笔者

在图 12-1 所示的大数据产业全景图中，老牌巨头无一缺席。微软、IBM、甲骨文、英特尔、思科、SAP、EMC 等公司，虽然它们的技术实力依然雄厚，但是业界的目光还是聚焦在几个新兴的巨头身上，它们是苹果、谷歌、亚马逊、Facebook。谷歌的前任 CEO 埃里克 • 施密特甚至为这四家公司提出“四大科技平台”的概念。毫无疑问，微软、IBM 等传统巨擘都被排除在外。为了行文方便，取这四家公司（苹果（Apple）、谷歌（Google）、亚马逊（Amazon）、Facebook）第一个英文字母组成缩写“FAGA”，选择倒序能让“FAGA”看起来更像一个单词。

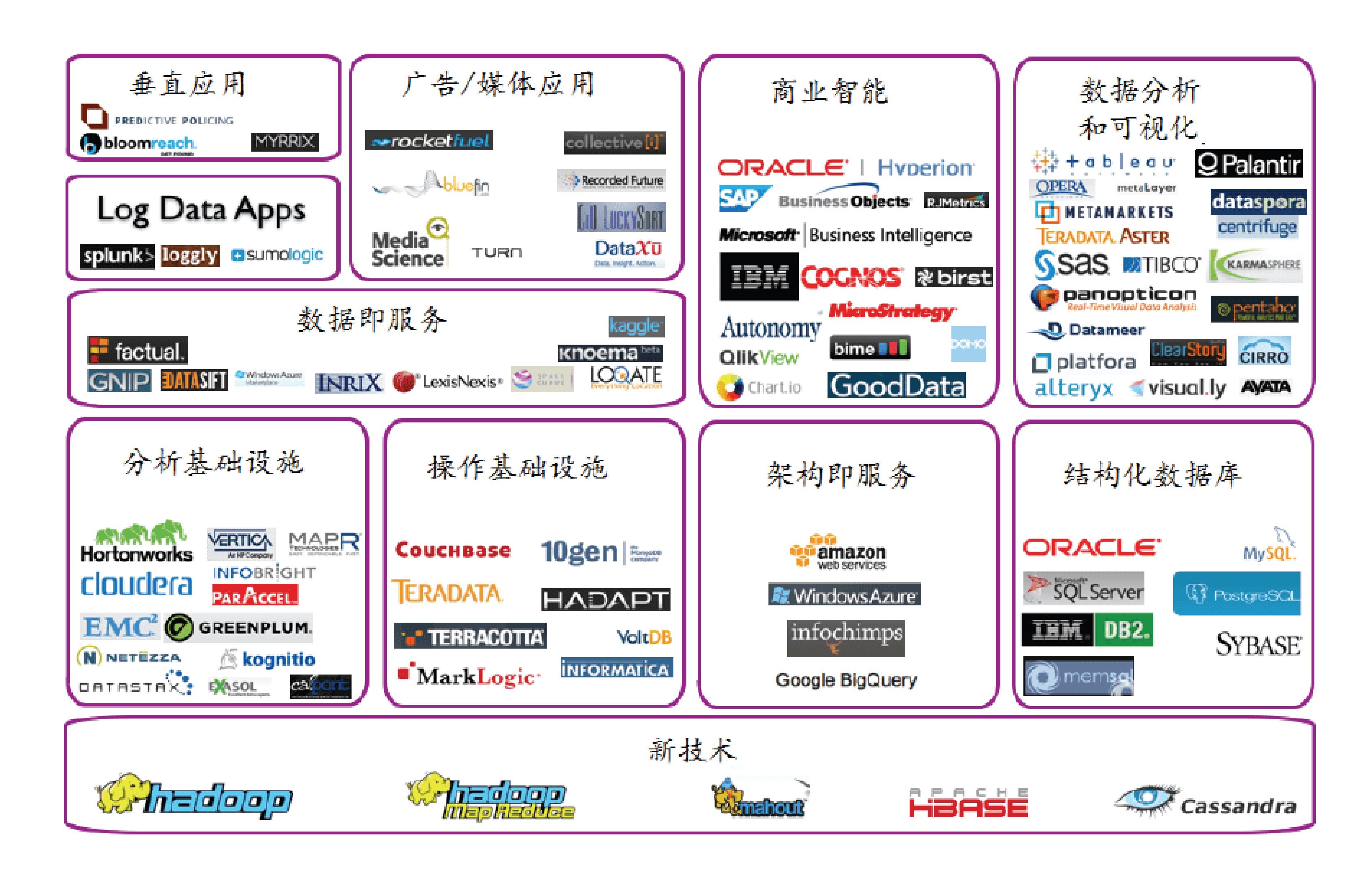


图 12-1　大数据产业全景图①

“FAGA”组合具备一个共同的特征，就是都有自己独一无二的“数据资产”，传统的巨头缺乏行之有效地收集数据资产的途径，而且数据运营似乎也不是它们的强项。另外，近十年引领信息产业发展的重要思想和技术创新，相当一部分来自“FAGA”组合。譬如，云计算是谷歌和亚马逊在 2006 年提出并付诸实践的，为公众所熟知却是苹果公司“iCloud”的功劳。更令传统巨头尴尬的是，“FAGA”组合都秉承“自己动手，丰衣足食”的理念，从底层芯片、服务器、操作系统到数据库都是自己开发，能不用就不用 IBM、甲骨文、微软等公司的商业化产品。

---

① 根据 Dave Feinleib 的原图翻译。出处：blogs.forbes.com/davefeinleib。

英特尔首席信息官（CIO）黛安•布赖恩特（Diane Bryant）几个月[①]前公布了一组有趣的数据。2008 年，75%的英特尔服务器芯片收入来自于三大服务器制造商，即 IBM、戴尔和惠普，然而到 2012 年，同样是 75%的收入，却来自八家公司，不再是上述三家巨头。设想，如果戴尔一年卖出 200 万台服务器，而某公司需求量为 100 万台，相当于戴尔一半的业务量，这时，这家公司便会意识到自己生产服务器的重要性。以谷歌为例，业内人士根据其庞大的数据中心耗电量来推测，谷歌大约拥有 100 万[②]台服务器。谷歌在英特尔列出的八大服务器制造商中，排名第五，但是这家搜索引擎巨头生产的服务器仅供自身业务需求使用，未对外出售。

新兴巨头“FAGA”与传统巨擘之间的博弈，天平朝 “FAGA”倾斜。因为它们既有庞大的数据资产，又具备处理技术，而传统巨擘空有一身本领，却缺少施展的舞台，如图 12-2 所示。再者，一些大型的商业用户喊出“去 OIE[③]”的口号，不愿意把自己珍贵的数据资产托付给昂贵的商业软件，以免出现被 IT 供应商“绑架”的局面。这个现象是“软件已死，数据永生”的内涵。

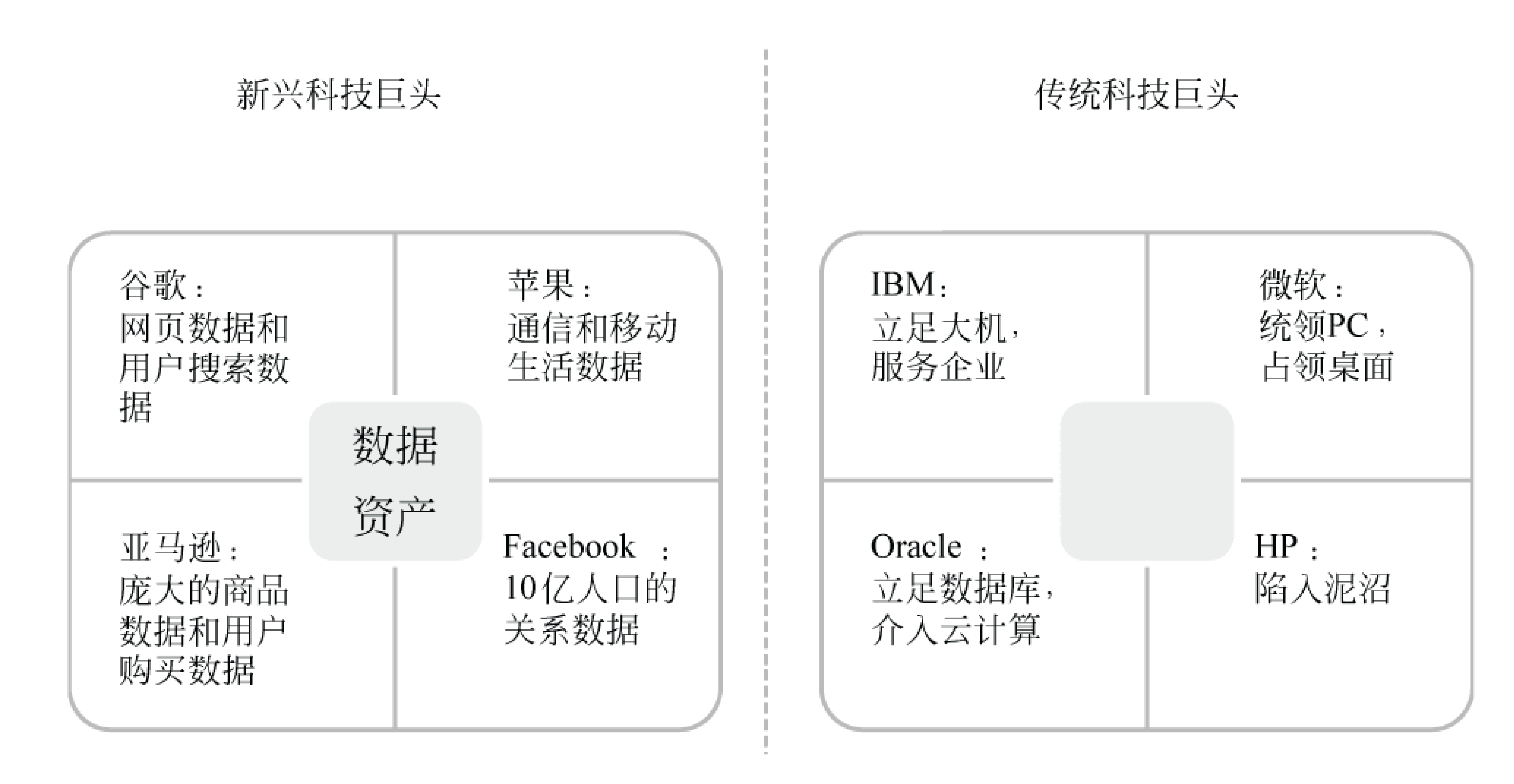


图 12-2　新兴科技巨头同时掌握数据资产和技术处理能力

① 指 2012 年。

② 据谷歌绿色能源团队项目经理大卫•雅各波维茨(David Jacobowitz)近日向美国斯坦福大学教授乔纳森•库米(Jonathan Koomey)提供的数据显示，谷歌数据中心所使用的电量不到 2010 年全球数据中心所使用的 1988 亿度的 1%，这意味着截至 2011 年年底谷歌拥有约 90 万台服务器。

③ OIE 是 Oracle（甲骨文）、IBM、EMC 三家公司名称的首字母。去 OIE，指一些大型机构开始自建 IT 基础设施，更多利用开源软件，逐步摆脱对商业软件、硬件的依赖。

第一节　传 统 巨 擘

本书第四章介绍行业变迁时，也谈到 IBM、甲骨文等公司的新产品和战略动向。本节只是概要说明这些公司过去赖以成名的独门绝技和未来发力的方向。它们依然实力强大，发展后劲十足，只是在大数据时代，不再夺人心魄而已。

蓝色巨人 IBM

IBM 公司的历史就是一部计算机发展史，单凭这一点，足见这家公司的传承和实力。因其蓝色的公司标识，人们常把这位计算机界的领导者称为“蓝色巨人”。IBM 公司成立于 1911 年，迄今已走过了 101 年的风风雨雨。根据 2012 年第三季度的数据，IBM 市值为 2300 亿美元。吴军博士在《浪潮之巅》一书中指出 IBM 成为科技界常青树的秘诀是“保守”。毫无疑问，保守使得 IBM 失去了无数发展机会，但也让它能专注于最重要的事，并因此立于不败之地。

IBM 一直致力于企业信息化市场，从不涉足个人消费业务。它的客户都是银行、电信、政府等大型行业或部门的龙头老大们。企业市场追求的首先是稳定可靠，从而使得 IBM 形成了保守之风。虽然远离个人消费市场，但是有两件事情令其声名远播，足以反映出 IBM 在数据处理和应用方面的深厚功力。

1996 年 2 月 10 日，IBM 公司的“深蓝”计算机首次挑战国际象棋世界冠军卡斯帕罗夫，以 2 比 4 落败。比赛在 2 月 17 日结束，其后研究小组把“深蓝”加以改良。1997 年 5 月 11 日，“深蓝”再度挑战卡斯帕罗夫，在前五局以 2.5 对 2.5 打平的情况下，卡斯帕罗夫在第六盘决胜局中仅走了 19 步就向“深蓝”拱手称臣。整场比赛进行了不到 1 小时，“深蓝”最终以 3.5 比 2.5 赢得了这场具有特殊意义的对抗，成为首个在标准比赛时限内击败国际象棋世界冠军的计算机系统。纵观“深蓝”的发展历程，IBM 研发小组向“深蓝”输入了 100 年来所有国际特级大师开局和残局的下法，并由美国特级大师本杰明将其对象棋的理解编成规则，进而由研发

人员编写程序教给“深蓝”。在和卡斯帕罗夫对弈时，每场对局结束后，工作人员都会根据卡斯帕罗夫的情况相应地修改特定的参数，“深蓝”虽不会思考，但这些工作实际上起到了强迫它学习的“作用”，这也是卡斯帕罗夫始终无法找到一个对付“深蓝”的有效办法的主要原因。“深蓝”的此次胜利，标志着人工智能技术又上了一个新台阶，人们从此将不得不认真地思考人与计算机的关系。从此，计算机在某些方面已足以与人较量。

在“深蓝”战胜卡斯帕罗夫之后 9 年，IBM 开始研制一款新型的超级计算机，这是一台以 IBM 创始人托马斯•沃森(Thomas Watson)命名的计算机——“沃森”，设计它的目的在于用纯自然语言来解答各种问题。在硬件方面，IBM Power 7 系列处理器是当时 RISC 架构中最强的处理器——采用 45nm 工艺打造的 Power 7 处理器拥有 8 个核心 32 个线程，主频最高可达 4.1GHz，二级缓存更是达到了 32MB。而在软件方面，IBM 研发团队为“沃森”开发的 100 多套算法可以在 3 秒内解析问题，检索数百万条信息然后再筛选还原成“答案”并输出成人类语言。

相比“深蓝”，“沃森”可以说在人工智能上前进了一大步。首先，它的计算能力大约是“深蓝”的 1000 倍；其次，国际象棋的规则定义非常明确，而人的自然语言完全是开放式的，往往很模糊，需要上下文才能理解意思；最后，在进行实际应用时，无需工作人员对各种参数进行调整，也无需连接互联网，因为它采用了大量的自然数据，而不是工程师输入的已知数据，完全依靠其 4TB 磁盘上的 2 亿页结构化和非结构化的信息进行判断，其领域知识库包括百科全书、字典、地理类和娱乐类的专题数据库、新闻报道、经典著作等。

2011 年 2 月，“沃森”参加美国最受欢迎的智力竞猜电视节目《危险边缘》，并击败该节目历史上最成功的两位选手肯•詹宁斯 (Ken Jennings) 和布拉德•鲁特 (Brad Rutter)，成为《危险边缘》节目新的王者。在比赛过程中，“沃森”在得到问题后，会进行一系列的计算，包括语法语义分析、对各个知识库进行搜索、提取备选答案、对备选答案证据的搜寻、对证据强度的计算和综合等等。它综合运用了

自然语言处理、知识表示与推理、机器学习等技术。它的主要技术原理是通过搜寻很多知识源，从多角度运用非常多的小算法，并对各种可能的答案进行综合判断和学习。这就使得系统依赖少数知识源或少数算法的脆弱性得到了极大地降低，从而大大提高了其性能。

在“沃森”参赛之前，它会从历史数据中进行学习。比如，如果它回答错了一个往期节目上的问题，它会从中学习到一些信息。在参赛之时，它主要依赖以前学习的结果，但也进行一些简单的在线学习。例如，它可以从已经被其他选手回答的同一类型问题中归纳出一些特点，指导自己回答这类问题。另外，答错题目也会导致“沃森”调整其游戏策略。因此可以说，“沃森”具备了初步自我学习和完善的能力。

据国外媒体报道，在“沃森”成功参加电视节目之后，IBM 日前已与美国俄亥俄州克利夫兰的一家医院签署了一份协议，计划将这台超级计算机投入于医生的培训工作当中，进一步为人类社会服务。

“人工智能”是大数据应用的高级阶段。总结人工智能从“深蓝”到“沃森”的发展历程，“通过机器的学习、大规模数据库、复杂的传感器和巧妙的算法，来完成分散的任务”是人工智能的最新定义。尤其是涉及机器学习、大规模并行计算、语义处理等领域，人工智能需要将这些技术整合在一个体系架构下以理解人类的智能和行为。蓝色巨人不经意间又走在了大数据的前沿。

微软的愁绪

PC 时代，微软是当仁不让的霸主。每当 PC 销量不振，一干厂商就把目光投向微软，期待它新一代的操作系统撬开消费者的荷包。比尔 • 盖茨也是最早提出平板电脑概念的人，早在 2002 年，微软就推出了一款“Tablet PC”的概念产品。尽管如此，移动时代的到来，还是让微软措手不及，成为起个大早赶个晚集的典型代表。微软眼睁睁地看着苹果和谷歌为了定义智能移动设备的产业生态和竞争格局打得不

亦乐乎，自己居然沦落为“打酱油”的角色。

Win8 铺天盖地的宣传之前，笔者就曾经写过一篇博客文章，看淡微软的前景，标题是“Win8，PC 时代的背影”，分析思路就是微软缺乏数据资产和运营能力。

只恐 Win8 舴艋舟，载不动微软的许多愁。

不看好微软平板 surface，也不看好诺基亚的 Win8 手机，并不是这些产品不好用，也不是缺少应用，而是在大数据时代，微软没有积累足够的竞争优势。微软依然生活在 PC 时代。看到比尔·盖茨介绍平板电脑的照片，两鬓斑白，不禁感叹，他已经老了。

曾几何时，盖茨每年的思考周被认为是保障微软核心能力的最佳实践；盖茨投资苹果 1 亿美元，估计再没有人想起。属于微软的时代已经远去，Win8 可能是留给 PC 时代的一个背影而已。

大数据时代，操作系统不再是产业链的中心，而被边缘化成渠道——吸引用户使用，成为采集用户行为数据的渠道。当数据变成了核心资产，微软构建的庞大软件帝国只能眼睁睁地成为看客，看着使用自家软件的用户数据源源不断地流到竞争对手的数据中心而无能为力。想一想谷歌、苹果、Facebook、亚马逊它们庞大的数据中心吧，这才是互联网企业的竞争壁垒。什么操作系统、平板、手机、阅读器等等都是浮云，表面上大家拼体验、拼配置，实际拼的是用户量和流量。

在大数据报告中，笔者指出衡量软件价值的标准：软件的价值和它带来的数据流量和活性成正比。如果分析微软的话，就把这个公式扩张一下，平板电脑的价值和它带来的数据流量和活性成正比。

回顾和微软竞争的一些重量级选手，凡是非互联网企业，都被微软霸权压迫的苟延残喘甚至消亡，直到谷歌异军突起，微软方乱了阵脚。

IBM 对微软是亦师亦友。盖茨敏锐地抓住为 IBM PC 开发操作系统的订单，开始了霸业之路。微软利用操作系统近乎垄断的优势，打压、收购各类应用软件公司。在 PC 时代，微软开创了售卖软件拷贝的商业模式，通过持续、不断地升级操作系

统，带动整个 PC 产业链的升级换代。只要 Intel 硬件计算能力提升，微软就发布庞大臃肿的软件，吸引、逼迫用户升级。当 PC 销售低迷的时候，各大 PC 制造商如惠普、联想、戴尔，就眼巴巴地盼着微软操作系统赶紧升级，带来新一轮用户换机高峰。

这个时代，微软是整个 PC 产业的核心，它的一举一动，无不牵动业界和用户的目光。大家一方面反对微软的霸权，一方面又不得不用它的软件，使用微软的软件，反过来又加强其霸权地位。在这种产业形态中，想要撼动微软几乎是不可能的事情。

IBM 看到养虎为患，最早出来打压微软。针对微软的 Windows 操作系统，IBM 开发了 OS2，寓意新一代的操作系统。估计现在好多人都没有听说过 OS2 的大名。这个系统具备现在苹果系统的某些特点，启动快、安全、不死机等等。单单考虑技术特性，这款系统无疑是强大的，但有它一个致命的缺陷，无法兼容 Windows 平台上海量的应用程序，如微软的办公软件。用户为了使用熟悉的应用软件，不得不选择忍受安全性差的 Windows。OS2 尽管有 IBM 这样的行业巨擘撑腰，也不得不很快败下阵来。

微软和 IBM 之争，可以得出的结论是：没有大量应用软件支持的操作系统，是没有前途的。

另一场惊心动魄的较量，在网景公司和微软之间展开。在互联网发展初期，网景开发出了首款易用的浏览器软件，风靡世界。微软奋起直追，利用垄断的操作系统，免费赠送 IE 浏览器。两家公司随即展开大战，网景有先发优势，微软有操作系统优势。曾经有笑话说，微软新版操作系统暂缓发布的原因为网景的浏览器运行得太顺畅了。还有一幅漫画调侃两家激烈的收购大战，画中一名大腹便便的人说：“第一步是开发产品，第二步是被微软或者网景收购。现在的问题是怎么绕过第一步。”在财大气粗的微软面前，网景公司逐渐感到力不从心，开始起诉微软不正当竞争。微软最终输掉了旷日持久的官司，但网景已经奄奄一息，最终被其他公司收购了。

谷歌创始人在回顾这段历史时说，网景没有利用浏览器优势转型成一家互联网公司，其经营模式依然是传统的卖软件拷贝的商业模式。这段公案得出的结论是：没有操作系统的优势，在工具软件领域挑战微软，是注定灭亡的。微软的一大竞争法宝就是和操作系统捆绑的策略。一旦微软祭出这件法宝，大多数公司都会俯首称臣，因为用户不会为了某一款应用软件而放弃操作系统。

当互联网时代来临，微软措手不及。

PC 时代，微软的商业模式是开发软件，用户付费购买软件拷贝的使用权。而在互联网时代，雅虎开创了对用户免费的互联网经济，谷歌更是把此模式发扬光大，开发了一系列软件供用户免费使用，却向广告主收费。谷歌的用户越多，它的搜索引擎就越精准，广告收入就越多。所以，谷歌的产品全部是免费的、在线的。这种彻底的互联网企业，真是革了微软的老命。微软的主要产品线，谷歌都有对应的免费产品。操作系统领域，微软卖 Windows，谷歌送安卓；办公软件领域，微软卖 Office，谷歌送 Docs。微软以前对付 IBM、网景等竞争对手的法宝，对谷歌完全失灵。微软的捆绑策略毕竟还是收费的，无非是用操作系统补贴应用软件，但谷歌是对用户彻头彻尾的免费。

当微软疲于应付谷歌时，苹果又异军突起，融合科技和人文特质的 iPod、iPhone、iPad，让微软相形见绌。通过这些卓越用户体验的终端产品，苹果积累了大量的注册用户，收集用户的喜好，如音乐、图书、游戏等等，拥有完备的支付通道——iTunes。苹果事实上是一家拥有互联网 DNA 的公司。苹果的护城河并不仅仅在于卓越的设计能力，其庞大的用户行为数据更是其所向披靡的法宝。用户可能会因为三星手机更炫而换掉手中的苹果，但不会丢掉 iCloud 中保存的照片和音乐。这些数据积累的越多，用户对苹果的依赖就越强。苹果所做的一切，都是为了吸引用户生成更多的内容，并提供更好的用户访问内容的体验。微软显然不具备这个能力，看看 MSN、Windows live 等在线应用的表现，就知道微软只能是 PC 时代的霸主了。

Win8 注定是 PC 时代的一个背影。

笔者相信 Win8 会很出色，拥有很好的触控体验，就像多年前 IBM 的 OS2，但是缺少应用而注定消亡。Win8 缺少用户内容，积累不了像谷歌、苹果一样的大数据，所以它注定是一个打酱油的角色。微软的财力可以保证持续地投入和研发，但这一切也只能保证 Win8 的存在感。Win8 的梦想很大，但其设想部分由苹果的 iOS 实现，部分会被安卓瓜分。它不过是 PC 时代的一个背影。

甲骨文的雄心

公司如人，甲骨文的创始人——拉里·埃里森是硅谷传奇中另外一位特立独行的人物，他热衷于滑翔飞机、帆船比赛等一些高风险的运动。有一则轶事可以反映埃里森的风格，当他听说竞争对手发明了一种新技术——分布式查询，十天后甲骨文就刊登广告宣布了 SQL 之星——第一个分布式查询数据库，而事实上当时没有任何这样的产品。埃里森就是这样，想象产品应该怎样，然后才去实现，如果成功了，他是成功的预言家，失败了，他就是骗子。但正是在拉里·埃里森的带领下，甲骨文公司的市值达到了 1500 亿美元，成为世界上仅次于微软的第二大软件公司。很难想象没有拉里·埃里森，甲骨文会怎么样。他在一次回答记者提问时说：“人生就是一场旅途。我们对他人和自己都非常好奇。这就是个探索极限的过程。我对科技相关的事物异常着迷。持续不断地探寻极限，学习用与他人竞争的方式来解决顾客问题。整个事情是如此的令人沉醉。我甚至不知道退休后还能做什么。当我扬帆远航时，我会环顾四周看看有没有人愿意比赛。我真的非常喜欢竞争。”

说到数据库，顺便谈谈甲骨文赖以成名的 Oracle 数据产品。

1970 年 6 月，IBM 公司的研究员埃德加·考特（Edgar Frank Codd）在 Communications of ACM 上发表了那篇著名的《大型共享数据库数据的关系模型》（《A Relational Model of Data for Large Shared Data Banks》）论文。这是数据

库发展史上的一个转折点，从这篇论文开始，关系型数据库软件革命的序幕被拉开了。

拉里·埃里森决定紧紧跟随 IBM 的步伐，开发通用的关系型数据库软件。甲骨文早期的数据库非常糟糕，但是它的销售力量强大，埃里森就是公司第一号的推销员。美国政府在关系型数据库推广过程中扮演了重要的角色，因为无论这些公司的软件有多差，政府总是会买单。很快甲骨文脱颖而出，成为数据库的领导厂商。在中国，电信业几乎全部业务应用都是基于甲骨文的数据库。IT 从业人员如果有甲骨文公司的认证，身价就会立即上升，甚至投标时，都要把有多少甲骨文认证的工程师作为中标的条件之一。

甲骨文是“大数据”的早期布道者之一，“大数据”是其未来的重点发展方向。关于甲骨文公司大数据产品方面的介绍以及产业垂直整合的介绍参见第六章，这里不再赘述。如果要在甲骨文、IBM、微软三者之间做选择的话，笔者更看好甲骨文。无他，只是因为拉里·埃里森——甲骨文的“永动机”。

第二节　新兴巨头

“FAGA”组合最近几年吸引了绝大多数人的目光，描写它们的书籍不胜枚举。本不需要笔者来画蛇添足，但是分析师的职业要求，必须对产业的发展做出预测，并且苹果、谷歌、亚马逊、Facebook 这四家公司之间的合纵连横，亦是非常精彩。所以，这里从数据资产角度来谈谈它们的发展方向：解释一下苹果为什么在自家地图应用尚未成熟的情况下，就悍然驱除谷歌的地图？为什么亚马逊一定要推 Kindle 阅读器？预测一下 Facebook 是否会选择做手机？为什么谷歌会拒绝为 Win8 开发应用？

在分析这些新兴巨头的游戏之前，来回顾第三章提到的“数据资产评估模型”，它们所做的一切都是围绕如何提升数据资产的规模、维度和活性开展的，继而在数据里面淘金。谷歌和亚马逊已经建立了数据淘金的完美商业模式，Facebook 面临的挑战是如何更好地变现 10 亿用户的“关系数据”，苹果主要收入来自移动终端的

销售，靠封闭的产业链打造了专属的后花园。这些新兴巨头最近推出的一系列软件服务，都是在利用数据淘金，意在成为凌驾于电信运营商之上的虚拟运营商。电信运营商则处在相对尴尬的角色，巨头都在向数据运营方向渗透，而它们却处于无险可守的窘境。

苹果市值一度超越了 6000 亿美元，是世界上最值钱的公司，但是未来谷歌和亚马逊都有问鼎全球“市值王”的潜力。不同的是亚马逊靠销售商品赚钱，而谷歌通过广告赚钱。它们两个的命脉都是通向它们网站的“流量”。越多的人访问谷歌，它的广告越值钱；越多的人访问亚马逊，它就会增加销售商品的机会。

谷歌终将超越苹果

苹果和谷歌的竞争是在定义未来的智能终端的产业生态。它们之间的竞争不仅仅是在智能手机上你的出货量大，我的出货量小这么简单。首先看看谷歌、苹果商业模式的差异。如图 12-3 所示，谷歌早早地建立了“大数据淘金术”，日进斗金，每天入账 1 亿美元。在第六章中已经详细介绍了谷歌搜索广告和内容广告的技术。

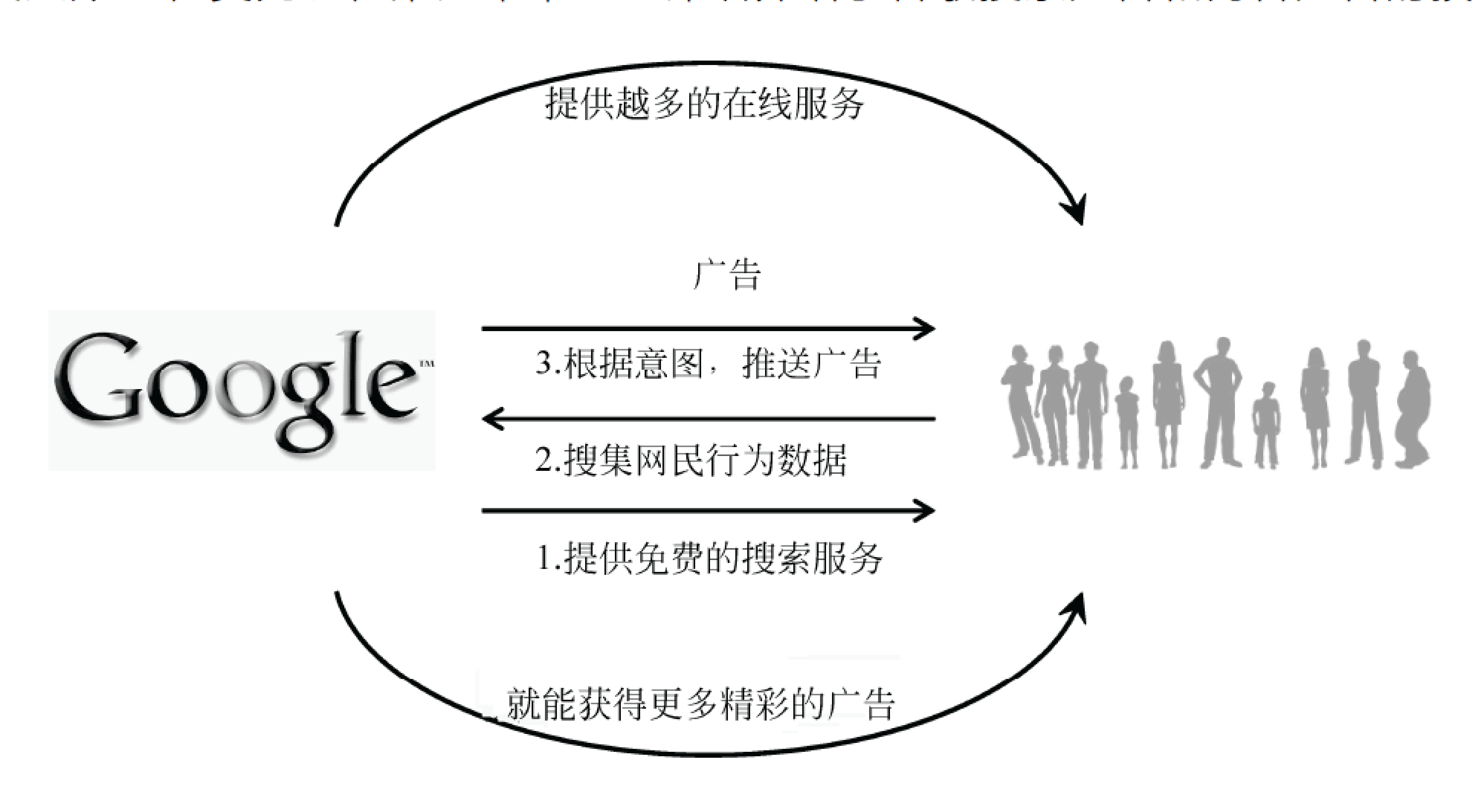


图 12-3　谷歌的炼金术

苹果公司的核心商业模式是销售各种智能终端，如以 iPhone 为代表的智能手机、以 iPad 为代表的平板电脑。苹果同样会提供各种各样的服务，但是苹果提供服

务的目的和谷歌略有差异，现阶段苹果提供的服务是为了让用户能在各个场合使用智能设备。譬如提供视频应用，可以让大家随时随地看电影，而不必端坐在电脑旁。苹果提供的各种应用越多，人们就会越喜欢越经常使用移动设备，从而促进苹果的销售。苹果公司的核心商业模式如图 12-4 所示。

所以，iPhone 问世之初，谷歌和苹果是相互协作、相互促进的。苹果设备上缺少吸引用户使用的应用，谷歌恰好可以弥补苹果的缺憾。于是两家开始了蜜月般的合作，甚至一度探讨公司合并的事宜。苹果设备卖得越多，谷歌产品的访问量就越大，谷歌就会获得更多的广告收入；同样，用户越是喜欢谷歌的应用，反过来也会带动苹果设备的销售。两家强弱互补，各取所需。

但是未来呢？未来属于谁？如果用户是因为喜欢谷歌的服务而购买设备的话，苹果设备的独特性就会丧失。更关键的是，谷歌和苹果的商业模式存在天然的深层次冲突。在谷歌的商业模式中，设备不过是收集用户行为数据的工具而已。对谷歌而言，什么样的设备不重要，重要的是使用这类设备的人群要足够广阔。只有更加广阔的人群，才会带来更多的数据、更多的流量、更多的广告客户。从谷歌的商业模式出发，谷歌的服务要占据各种各样的设备，从台式机到手机、平板电脑，甚至是智能电视、智能眼镜、智能汽车等等，而且这些设备要尽量廉价。

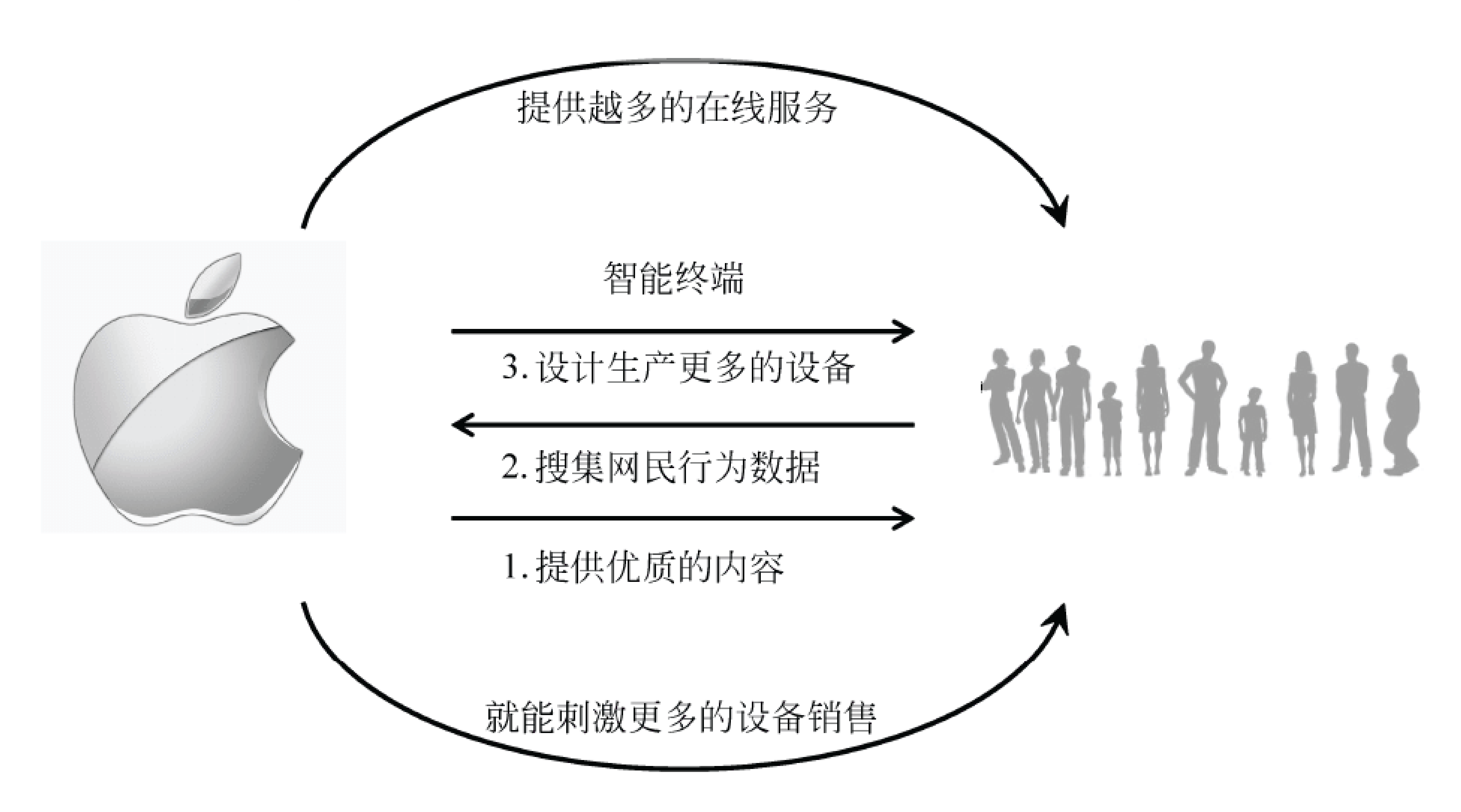


图 12-4　苹果的核心商业模式

对苹果而言，恰恰相反。如果按照谷歌的思路，苹果的设备势必越卖越便宜，

而且卖这些便宜的设备都是为谷歌做了嫁衣：辛辛苦苦圈来的用户，最后都变成谷歌的数据资产，势必被谷歌所掌控，这是苹果坚决不能允许的。而谷歌面临的困境是，苹果掌控着终端，随时可以驱逐谷歌的应用。是委曲求全，还是另觅出路，这是谷歌不得不考虑的问题。

最终世界都看到了谷歌的选择，推出开源、免费的 Android 操作系统，授权智能终端制造商生产谷歌可以掌控的设备。所谓谷歌掌控，是指那些使用 Android 系统的智能终端，都必须搭载谷歌的服务，如邮件、地图、音乐、搜索、股票、旅游等。2012 年闹得沸沸扬扬的“阿里云”手机事件，就是阿里试图推出去谷歌服务的 Android 系统，被谷歌果断打压。

从此在智能设备领域形成两大阵营，谷歌对决苹果。这两大巨人有一个目标是相同的，就是促使销售更多的智能设备，而非仅仅是智能手机。因此战火蔓延到整个智能终端市场，顺便挤占了另外一个巨头的生存空间，那就是微软——昔日 PC 的霸主。

微软的创始人比尔•盖茨早早预测到了智能终端的影响，甚至十年前就开始研发样机，可惜的是，并没有在这场决定产业格局的战役中获得先发优势。他过度地依赖 PC，试图以个人计算机为中心扩展微软帝国，但不料想被苹果的平板电脑打了一个措手不及。当微软试图重振雄风时，产业格局的竞争已经超越智能手机层面，变成完整的智能设备生态环境的竞争，如图 12-5 所示。

未来是多屏融合，杀手锏是不同屏幕上一致的体验，如图 12-6 所示。举例而言，当笔者开车去一个陌生的体育馆看比赛时，首先可能会在智能手机上搜索到体育馆的位置，发动汽车的同时，车载电脑就会获取到智能手机上的导航信息，开始自动导航。如果看了一半回家，打开电视，就能在笔者离开的时间点继续观看转播的比赛。这就是谷歌、苹果现在竞争的重点，目前战火正旺的是智能手机、平板电脑，很可能到本书出版的时候，它们就会在智能电视领域打得不可开交了，而未来三五年，汽车或许成为另外一个战场。

微软在智能手机竞争时代已经落后了，在多屏融合的时代正发力急追。Win8 就是承载了微软梦想的产品，但是因为 Win8 没有经过智能手机竞争的洗礼，没有积累足够的用户数据，缺少数量众多的应用，所以在多屏融合的竞争中也处在不利的位置。



图 12-5　智能终端的产业生态，竞争要素再次发生了变化[①]

	2005	2010	2015
愿景	三网融合	一部手机 数万应用	一个生态 多个屏幕
焦点	网络	设备	生态系统
竞争基础	服务价格	应用数量	一致的用户体验

图 12-6　多种设备之间，一致的用户体验成为竞争的重点[②]

① 出处：VisionMobile。
② 出处：VisionMobile。

谷歌显然不愿意把自己的服务加载到 Win8 上。Win8 为谷歌带来的流量微乎其微，但是谷歌应用却可以帮助 Win8 的销售。如果谷歌没有自己的 Android 系统，为了制衡苹果，一定会帮助 Win8 的成长。但是现今而言，谷歌为 Win8 开发应用，显然不利于自己 Android 系统的销售。因此，可怜的 Win8 只好靠微软的财力苦苦支撑。好在微软这个大财主有钱，所以一定可以维持 Win8 的发展，但是短期内不要指望有任何起色。

从商业模式来看，谷歌依赖广告的模式更加健康，因为无论是用户还是广告客户都会在谷歌的商业模式中持续受益。谷歌将不遗余力地推进智能终端的普及，并且是在尽可能低的价格上提供尽可能好的服务。历史上，凡是提供质优价廉服务的公司，几乎没有失败的先例。谷歌通过累积数据资产，挖掘数据价值，几乎完美地践行了这条商业定律，因此必将得到用户的支持。

尽管笔者非常喜欢苹果的产品，但是像“重新发明手机”这样的创新，不是年年都能发生的故事。笔者期盼苹果持续创新，引领潮流。但如果现在让笔者在谷歌、苹果、微软三者之间下注的话，笔者押宝谷歌。原因很简单，谷歌对数据资产的积累和运用远远超过了苹果和微软。

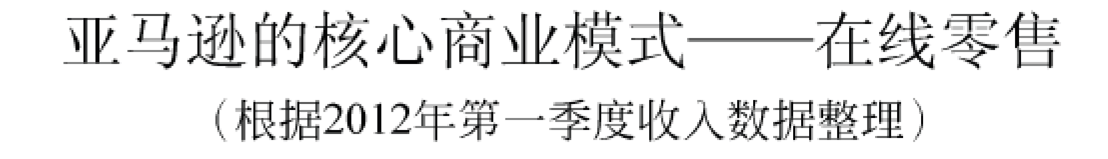
市场统计数据也证实了谷歌的实力，根据 IDC 的统计数据，2012 年第三季度，装有 Android 系统的手机全球销量是 1.36 亿部，占据 75%以上的市场份额。而三星也成为世界上出货量最大的手机厂商，终结了诺基亚长达 14 年的“最大手机厂商”的名头。众所周知，三星手机大量采用的正是谷歌的 Android 系统。

亚马逊亦可问鼎全球第一市值的宝座

亚马逊的成功，源自杰夫 • 贝索斯的远见。1997 年贝索斯发表了一封给股东的公开信，后来的公司年报中反复提及这封信，如果有人质疑亚马逊的商业模式，贝索斯就提醒投资人去读一读 1997 年的这封信。信中重点强调“It’s all about a long term”,反复提醒投资者，这是长期的生意。这封信的发表已经距今 15 年了，15 年

间，亚马逊已经从 1997 年 1.5 亿美元的销售额，暴涨到 2011 年的 480 亿美元，跻身千亿美元市值俱乐部。

IT 业界提到亚马逊，大多谈它的 AWS[①]，即亚马逊网络服务，为中小企业提供云计算的基础服务。据美国调查公司 451Group 的报告，AWS 已经占据了美国 59% 的云计算基础设施及服务（IaaS）市场份额，领先优势相当明显。尽管如此，亚马逊依然是一家不折不扣的在线零售商，如图 12-7 所示。



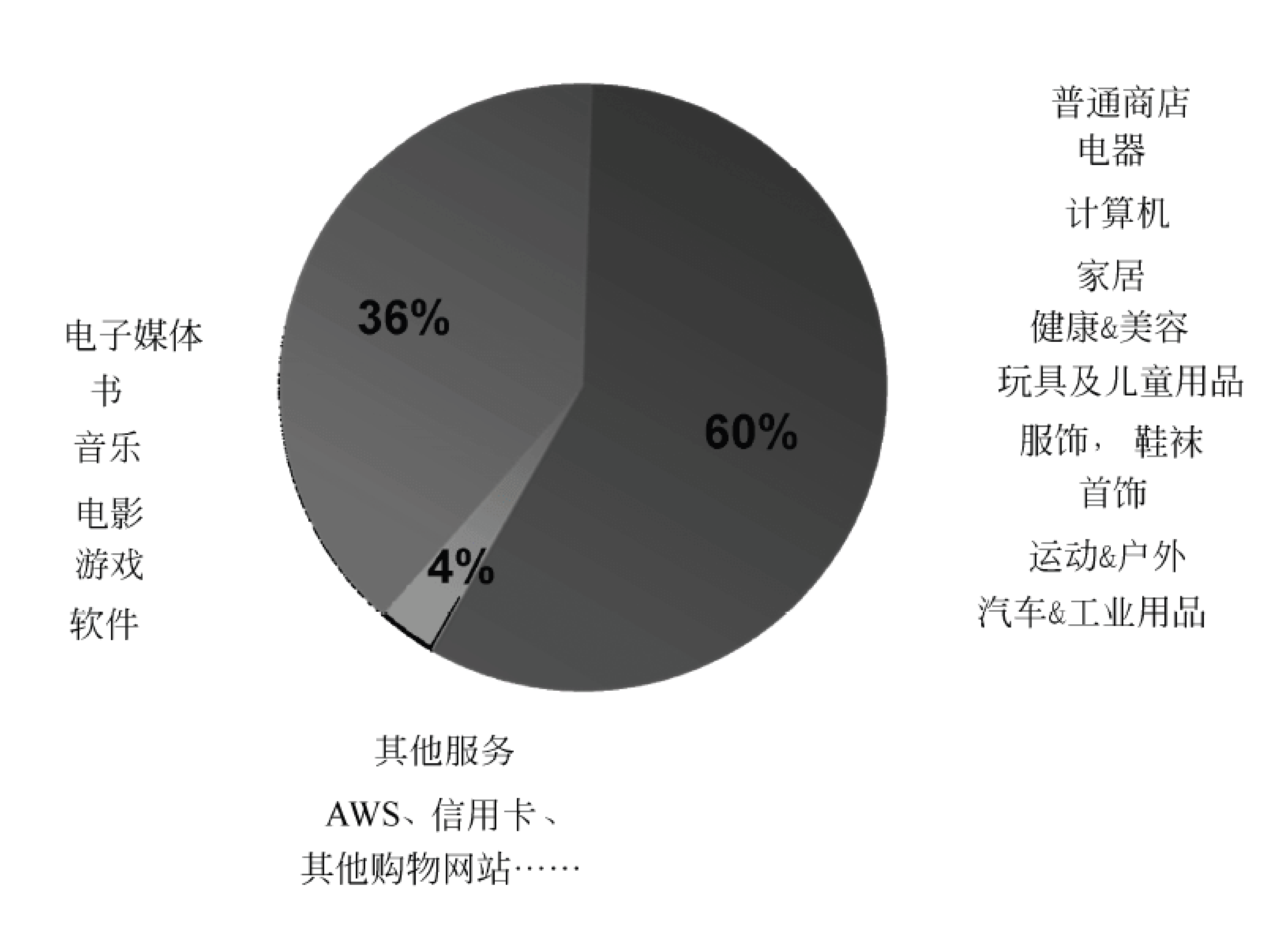


图 12-7　亚马逊的主要收入依然来自于销售普通商品

详细介绍亚马逊超出了本书的主旨，本书重点还是集中在亚马逊的商业模式，以及它如何利用数据资产来和谷歌、苹果展开竞争上。如图 12-8 所示，亚马逊作

① 亚马逊网络服务（Amazon Web Services）为亚马逊的开发客户提供基于其自有后端技术平台，通过互联网提供的基础架构服务。利用该技术平台，开发人员可以实现几乎所有类型的业务。亚马逊网络服务所提供服务的案例包括亚马逊弹性云计算（Amazon EC2）、亚马逊简单存储服务（Amazon S3）、亚马逊简单数据库（Amazon SimpleDB）、亚马逊简单队列服务（Amazon Simple Queue Service）、亚马逊灵活支付服务（Amazon FPS）、亚马逊土耳其机器人（Amazon Mechanical Turk）以及 Amazon CloudFront。

为世界上最大的在线商店，它的商业模式简洁明了[①]，就是吸引更多的流量[②]，卖出更多的商品。所以哪里能带来流量，亚马逊的广告或者服务就会出现在哪里。

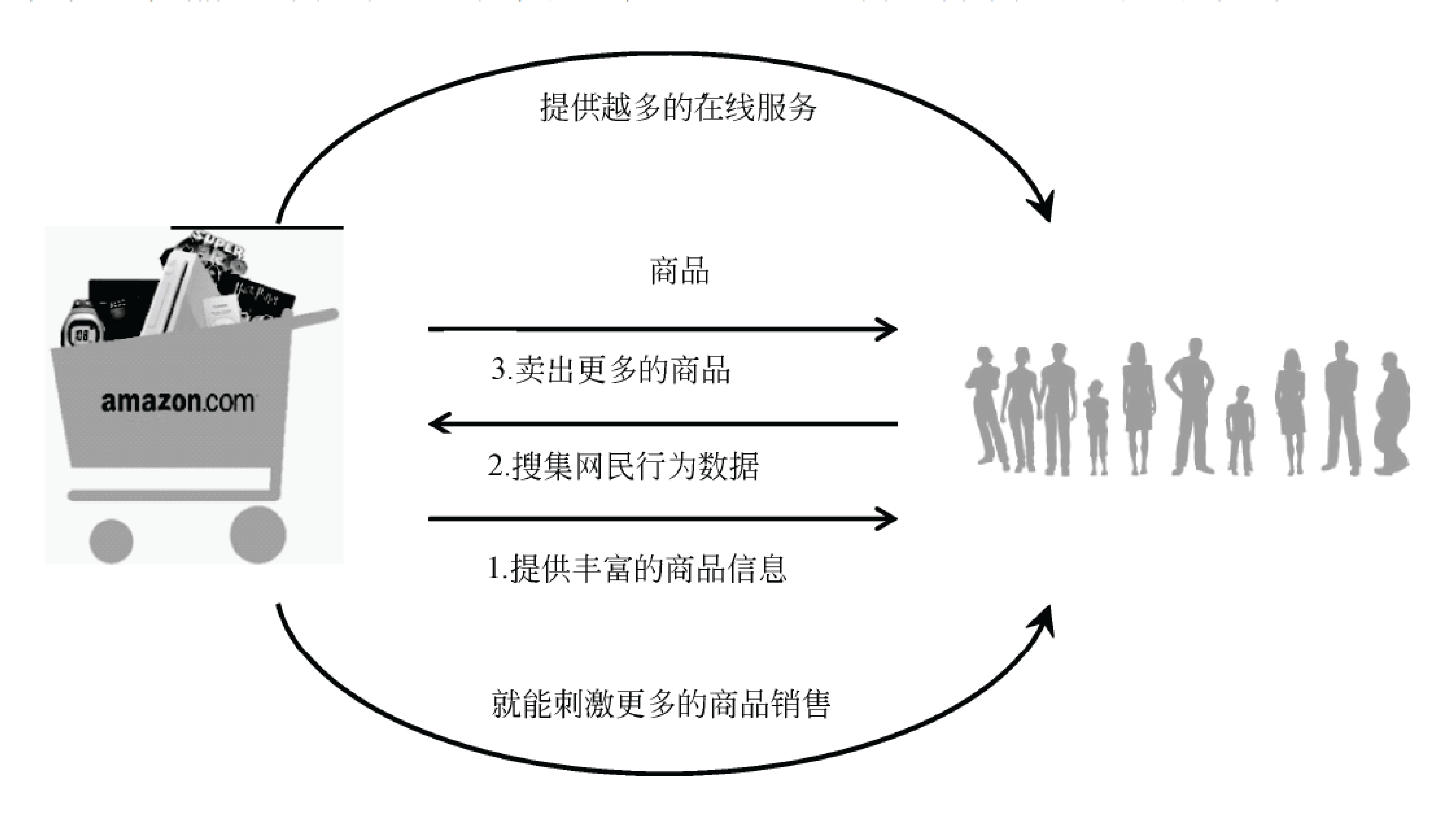


图 12-8　亚马逊的所作所为都是为了卖更多的商品

亚马逊也是一家不折不扣的大数据公司，其对数据资产的运用能力甚至与谷歌也不遑多让。亚马逊的 S3 云存储平台能够为地球上的每个人存 82 本书。和谷歌不同的是，亚马逊有庞大的“非数据资产”——商品库房，2011 年其拥有 50 个 2500 万平方米的库房，相当于 700 个麦迪逊广场。

根据亚马逊的统计，访问亚马逊网站的用户中只有 16%的人有明确的购买意图。也就是说，“逛街”的居多。如何让“逛”亚马逊“街”的用户“下单”是其核心的竞争力。亚马逊的大数据资产和大数据挖掘技术就是在这个环节派上用场的。它的“推荐系统”非常出色，大约有 20%～30%的订单都是推荐系统促成的。在本

① 事实上，电子商务已经发展出各种类型，中国的电子商务市场亦非常发达。本书并非专门研究电子商务业态，所以把其商业模式做了抽象的整理，利用其说明本书的主旨即可。

② 流量类似商店的客流，人流越多的地方，商机越旺。在互联网上，流量常常反映为网页的点击量、停留时间等指标。

书第六章中把亚马逊的推荐系统归类为“行为广告”的典型代表。

苹果 iPad 问世后鲜有敌手，只有亚马逊的 Kindle 系列平板电脑独树一帜和 iPad 抗衡。个中缘由正是亚马逊庞大的数据资产在发挥作用。根据 IDC 的统计，2011 年第三季度 iPad 占据平板电脑将近 62%的市场份额，其他杂七杂八加起来大约不到 39%。但是亚马逊发布 Kindle Fire 之后，短短的一个季度就占领了 17%的市场份额，成为和 iPad 分庭抗礼的潜在对手，如图 12-9 所示。

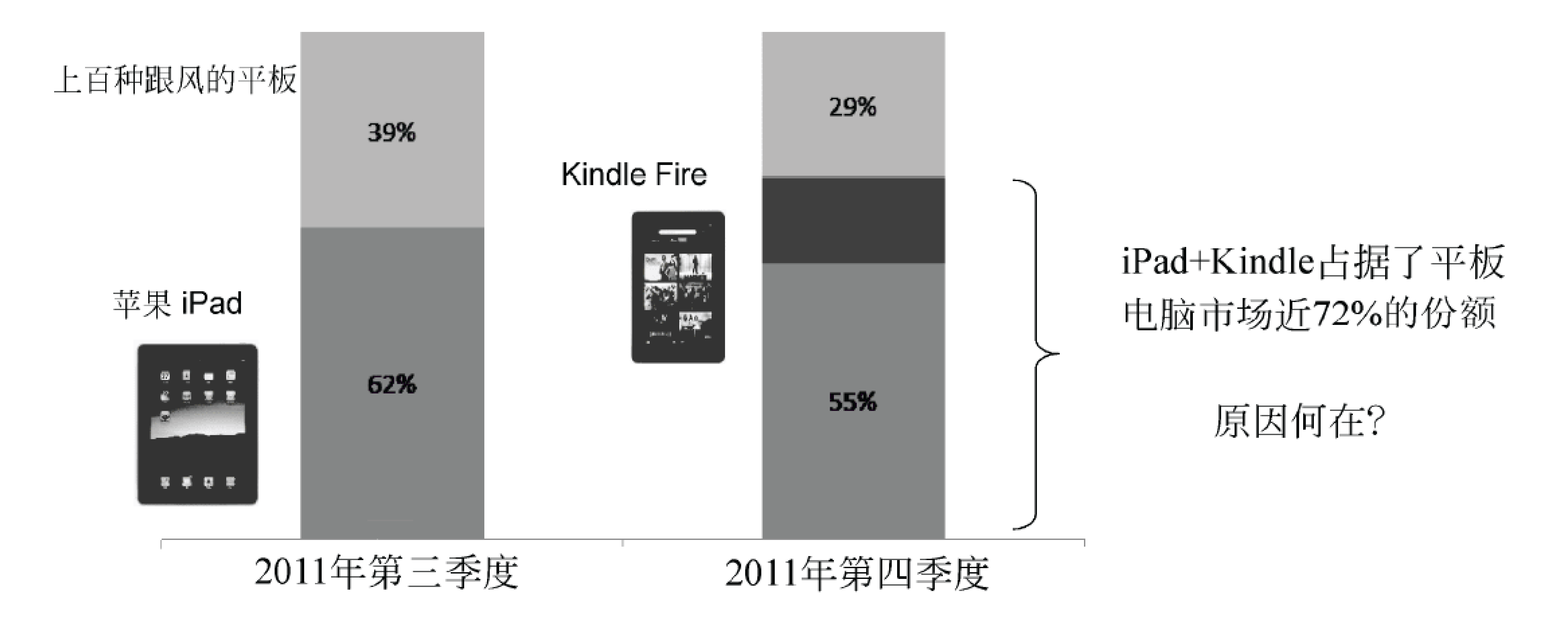


图 12-9　亚马逊的 Kindle 是目前唯一可以抗衡苹果 iPad 的平板电脑①

iPad 和 Kindle Fire 总共占据了近 72%的市场空间，进一步挤压了其他平板电脑的生存空间。事实上，亚马逊 2011 年以 79 美元，大约每台亏损 5 美元的价格销售 Kindle，一方面亚马逊有此雄厚的财力，另一方面 Kindle 成为人们访问亚马逊网站，进一步购买书籍、电影、音乐等电子内容的有效载体。亚马逊销售 Kindle，醉翁之意不在酒，更多的是要给自己的网站带来更多的流量，刺激用户购买相关的内容产品。如图 12-10 所示，其他 iPad 的跟风者，既没有亚马逊海量的内容数据库，又缺少 iPad 流畅的操作体验，只好沦落到低价竞争的境地。

① 数据来源：IDC。

	核心业务	附加业务	用户主要使用场景
iPad	设备	零售	数字生活设备
Kindle Fire	零售	设备	消费数字化内容
其 他	设备	—	模仿iPad

图 12-10　Kindle Fire 与 iPad 定位不同

根据大数据时代三大发展趋势来判断，亚马逊未来很可能会干以下三件事情，每件事情都可以引起产业的震动。

第一，为了进一步把控“流量”，亚马逊很可能推出自有品牌的智能手机。媒体上关于这件事情炒作得沸沸扬扬、真真假假。但是仔细推敲“行业垂直整合”的趋势，亚马逊智能手机是不得不为的事情。否则，很可能会被谷歌在上游“劫持”流量。

第二，购买亚马逊商品，可能会赠送上网的流量费。现在如果买 Kindle，是赠送包月流量的。未来这个模式有可能扩展到其他服务，如买到多少商品，就免全月的上网费等。也就是说，亚马逊有可能介入基础电信运营领域，就像谷歌正在做的事情一样。

第三，扩展旗下的小额贷款业务，就像中国的阿里巴巴正在做的事情一样，创新的金融服务将是亚马逊大力拓展的业务。

综上所述，亚马逊是一家极具张力的公司，它的业务疆域远远没有终结，凭借其独一无二的庞大数据资源和研发实力，亚马逊将是苹果和谷歌的强劲对手，这三家公司之间的争夺，不仅仅会影响信息产业的走势，势必波及电信运营、金融、零售、出版、教育、制造等各行各业。

Facebook 要抄谷歌的后路

Facebook 在 2012 年上市之初缔造了又一个资本神话，一度超越千亿美元市

值大关，随后大幅跌落，现在反弹稳定在 700 亿美元左右。这家公司的月度活跃用户已经突破 10 亿大关，几乎接近整个中国的人口，这是其庞大的独一无二的数据资产，如图 12-11 所示。

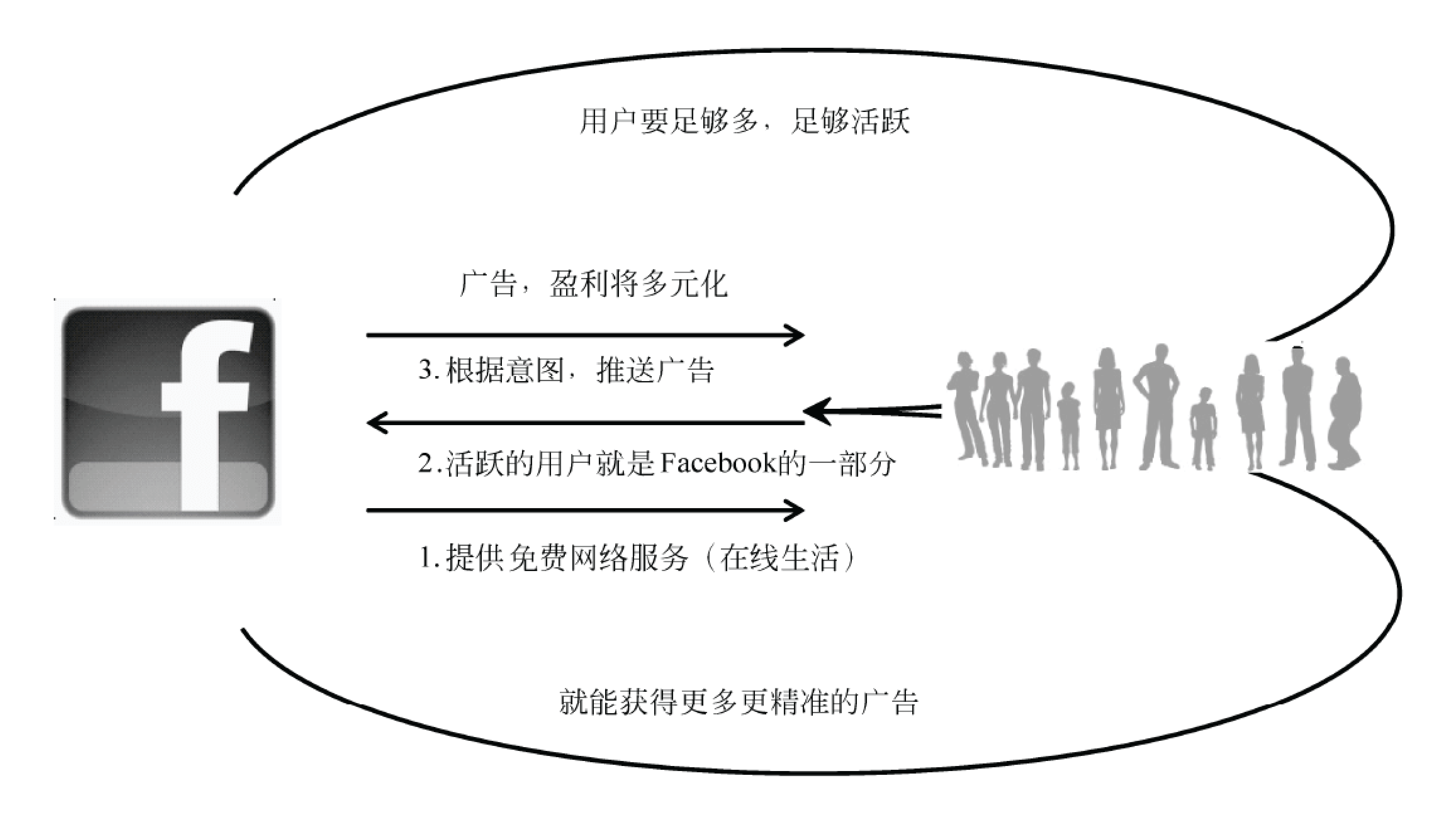


图 12-11　Facebook 的用户就是 Facebook 的一部分

当谷歌一骑绝尘领跑搜索广告市场之际，许多公司都在奋起直追。雅虎中断了与谷歌的合作，推出自己的搜索引擎。微软推出“Bing”搜索引擎，中文名称译为“必应”，有点有求必应的意思。从多年的市场表现来看，尽管谷歌的直接竞争对手们财大势雄，但却没有一家具备撼动谷歌根基的能力。

当大家的资本、技术、管理水平相差无几的时候，谷歌凭借先发优势积累的数据资产，成为后来者难以逾越的门槛。谷歌无时无刻不在编制互联网的索引。有资料显示，谷歌发现一个新建立的网页，只需要 4 小时。换句话说，利用谷歌的搜索引擎，世界离我们指尖只有 4 小时的“距离”。而且这个庞大的互联网索引，根据每个人的搜索请求还在不断的优化。越多的人使用谷歌，谷歌的技术就会越趋近完美。

它建立起来一个完美的“飞轮”，众多的搜索用户在推动这个飞轮不断地加速运转，后来者只能望其项背。这也是微软虽有庞大的研发队伍，富可敌国的资金也奈何不了谷歌的本质原因。

但是 Facebook 横空出世，搅乱了谷歌的步伐。Facebook 自身拥有 10 亿用户，几乎每天都在更新自己的所见所闻，喜怒哀乐。Facebook 可以说是这 10 亿用户自己一点一滴、每分每秒，聚沙成塔、集腋成裘建立起来的独立王国。这个“王国”的数据，谷歌没有办法得到。无论谷歌的搜索引擎算法多么高明，Facebook 之于谷歌就像“黑洞”般的存在。人们如果查询“好朋友们都喜欢去哪些餐馆？”诸如此类的问题，谷歌是无能为力的，而 Facebook 却易如反掌。

Facebook 在 2013 年 1 月 16 日，隆重地发布了“Graph Search”服务。Facebook 官方解释发布这个搜索引擎的初衷：“人们用搜索引擎去寻找答案。但是我们可以回答一系列没有人能够回答的问题，所有的这些都来源于 Facebook 本身的社交数据，而除了 Facebook 这些数据是无从得知的。因为这是人们分享和关心的数据。除了将人们联系起来，没有别的方法可以了解到人们所关心和分享的事物，满足人类对于发现的需求。所以，我们决定做搜索，因为只有我们可以这样做搜索。”

IT 业的魅力正源于此。回顾历史，我们发现每当一个重大的技术革新，就会诞生新的巨人。大机时代，IBM 号称蓝色巨人，后期 100%垄断大机的生产和服务。微机时代，微软横空出世，左手操作系统，右手办公软件，双剑合璧，天下无敌，不知多少公司在微软的凌厉攻击下，灰飞烟灭。互联网时代，微软眼睁睁看着谷歌迅速成长，虽然极尽扼杀之能事，却苦于无法匹敌谷歌的数据资产和商业模式，终于接受了谷歌做大的现实。同样，谷歌应对 Facebook 的挑战，亦无良策。Facebook 上用户自身即构成了 Facebook 庞大数据资产的一部分，谷歌只能望洋兴叹，退而求其次，自己退出“Google+”社交网络服务。但在用户的规模和活跃度上，谷歌与 Facebook 存在数量级的差距。数据资产不同，注定谷歌与 Facebook 拥有不一样的未来。

导读：

1. 这些新兴的公司就像在一片大数据沙漠中淘金的片片绿洲，充满了勃勃生机。正是有它们义无反顾的实践，才有可能拉开大数据时代的帷幕。正是对它们的系统梳理，我们才深深体察到产业的脉动、变化节奏和未来方向。
2. 这些公司依然属于狭义信息科技领域，分布在信息基础设施、数据分析基础设施、数据库、商业智能、广告/传媒、数据即服务、垂直应用等子领域。遗憾的是笔者难以穷尽所有的公司，选择其中有代表性的以飨读者。
3. 事实上，笔者更加关注传统行业与大数据对接的领域。传统行业在生产经营中同样积累了丰富数据，这些犹如沉睡的宝藏等待人们发掘利用。笔者反复强调大数据思维，就是试图让更多的行业了解到数据的价值，吸引更多的优秀人才投身到大数据中去，使传统行业获得更高、更快的增长。笔者将密切关注这方面的进展。

第十三章 创新凶猛

他们正在重新定义未来！

——笔者

“近一百多年来，总有一些公司很幸运地、有意识或无意识地站在技术革命的浪尖之上。AT&T、IBM、Apple、Intel、Microsoft、Cisio、Yahoo、Google、Facebook都先后被幸运地推到了浪尖……在这十几年到几十年间，它们代表着科技的浪潮，直到下一波浪潮的来临。……对于一个人来讲，一生赶上这样一次浪潮就足够了。对于一个弄潮的年轻人来讲，最幸运的莫过于赶上一波大潮……”[①]

2012年，国内外数据服务领域的投资案例数在所有投资行业中名列前茅。从“大数据”概念的刚刚提出，到现在仅仅几年时间，大数据相关的创业公司便如雨后春笋般冒出，并受到了产业界、学术界和资本市场的极大关注和热捧。这一现象表明，“大数据”已不是空泛的概念，而是以非常快的速度在各个领域落地生根。“大数据”已经开始逐渐兴起并形成一波新的浪潮，而这一浪潮在未来将会进一步得到强化，并推动传统产业的升级。

现有新兴已融资的数据服务类公司，服务的覆盖领域已非常广泛，从 NoSQL数据库、操作基础设施、数据分析、商业智能、广告/媒体应用到各个细分垂直领域的应用等等，笔者梳理出了以下典型的数据服务领域的创业企业，它们正在以大数据驱动传统产业，让传统产业变得更有智慧、更有洞见。

篇幅所限，笔者无法一一列举所有的新兴公司，仅仅从不同的服务领域，选取有代表性的公司介绍给大家，这些公司都在大洋彼岸。笔者同时了解到国内亦有部分公司，早早开始大数据领域的实践，苦于相知之日短，无法在本版中呈现给大家。成书之际，又大幅删减了公司产点介绍和客户情况。如需详细信息可以和笔者联系大数据主要新兴公司一览表见表13-1。

表13-1 大数据主要新兴公司一览表

大数据新兴公司	服务领域
社交数据平台 DataSift	数据即服务
开放平台数据提供商 Junar	数据即服务
数据分析平台 Precog	数据分析

① 摘自《浪潮之巅》的前言。

续表

大数据新兴公司	服务领域
分布式文档存储数据库提供商 10gen	操作基础设施
企业云存储服务 Nirvanix	操作基础设施
非结构化数据库解决方案 Clustrix	操作基础设施
大数据高效管理 RainStor	操作基础设施
用户行为监测分析 Mixpanel	广告/媒体应用
商务数据分析决策 SumAll	商业智能
敏捷数据管理 Delphix	商业智能
数据分析决策集成服务 GoodData	商业智能
数据智能解决方案 Ngdata	商业智能
智能搜索引擎 Attivio	商业智能
实时大数据分析 ParStream	分析基础设施
云笔记存储服务 Evernote	垂直应用
内部销售和线索反馈管理 InsideSales	垂直应用
职业搜索引擎 TalentBin	垂直应用
医疗保健行业大数据解决方案 Predilytics	垂直应用
大数据分析应用 Datameer	数据分析和可视化
跨平台大数据处理 Trifacta	数据分析和可视化
数据库服务商 DataStax	数据库

第一节　数据即服务

社交数据平台 DataSift

社交数据平台 DataSift 帮助“开发商以及第三方”访问 Twitter，Facebook 及其他社交数据资源，DataSift 能够对海量社交数据进行分析，向品牌公司、传统企业、金融市场、新闻机构等提供实时的或者历史的社交数据。

DataSift 创建于英国，但之后很快就搬到了旧金山。DataSift 于 2010 年 1 月

获得 150 万美元种子基金;2011 年 7 月收到 A 轮融资 600 万美元,GRP Partners 以及 IA Ventures 投资;2012 年 5 月获得后续 A 轮融资 720 万美元，由之前的 GRP Partners 以及 IA Ventures 投资；2012 年 11 月，获得 B 轮融资 1500 万美元，B 轮融资由 Scale Venture Partners 领投，Northgate Capital 和 Daher Capital 跟投。

DataSift 网站：http://datasift.com。

开放平台数据提供商 Junar

Junar 致力于成为开放数据后的专业处理引擎，从而为世界顶级公司和上百万的数据用户提供技术支持。如今我们正在快速迈入一个以信息和服务为主导的转型社会中，Junar 致力于推进这种转型并且推进数据经济的发展。

Junar 通过发布领先的基于云的开放数据平台加强了新兴数据经济。平台使得商业、政府和其他机构释放他们的数据以驱动新的机会、合作和透明性。通过 Junar，用户能轻松地选择收集什么数据、决定如何呈现数据，以及什么时候应该发布数据。客户能决定哪些数据集对公众可用以及哪些数据集只能内部使用。它被视为分享信息的下一代数据管理平台。

2012 年 9 月，Junar 获种子 A 轮融资 120 万美元，投资方包括 Aurus、Austral Capital，以及一群来自美国和拉丁美洲的天使投资。

第二节　操作基础设施

分布式文档存储数据库提供商 10gen

10gen 是 MongoDB（分布式文档存储数据库）技术的开源开发商，主要通过为客户提供支持、培训以及相关咨询服务获得收益。

10gen 成立于 2007 年，由在线广告公司 DoubleClick 前创始人兼首席技术官德怀特梅里曼（Dwight Merriman）和 ShopWiki 创始人兼首席技术官艾略特　霍洛维茨（Eliot Horowitz）共同创办。因此可以说 10gen 是一家由两位 CTO 工程师创办的企业。2011 年 9 月 10gen 获得由红杉资本（Sequoia Capital）牵头，Flybridge Capital 及 Union Square Ventures 两家风投共同完成的新一轮融资，融资额度高达 2000 万美元，融资总额达到 3100 万美元。2012 年 5 月 10gen 获得由 New Enterprise Associates 领投的新一轮 4200 万美元投资，其他投资方还包括 Sequoia Capital，Flybridge Capital Partners 以及 Union Square Ventures。截止这轮投资，10gen 已经获得总计 7300 万美元的投资，其估值达到了 5~5.5 亿美元。

10gen 网站：http://www.10gen.com/。

企业云存储服务 Nirvanix

Nirvanix 是一家企业级云存储服务商，公司发展速度很快，2012 年第一季度的营收几乎相当于 2011 年全年的收入。CEO 斯科特 • 杰纳洛克斯指出 Nirvanix2011 年管理的数据流量相当于前三年数据流量的总和，现在公司完成了世界上最大的私有云架构。

Nirvanix 公司于 2007 年 9 月份创建，公司从几个技术投资商那里总共获得了 1800 万美元的投资，投资商名单中包括了大名鼎鼎的 Intel Capital。Nirvanix 存储网络已经出现在遍布全球的很多数据中心之中。2007 年 9 月底，Nirvanix 又获得了 1200 万美元的融资，据悉这笔资金将用于 Storage Delivery Service（SDS）业务的开展，此轮融资是由风险投资机构 Mission Ventures 和 Valhalla Partners 主投，Windward Ventures 等跟投。2009 年 B 轮融资 2800 万美元，公司全力进军企业云存储。2012 年 5 月，公司第三轮融资 2500 万美元，资本总额增至 7000

万美元。此轮融资由风投公司 Khosla Ventures 牵头，Valhalla Partners、Intel Capital、Mission Ventures 以及 Windward Ventures 等投资公司共同参与。

Nirvanix 网站：www.nirvanix.com。

非结构化数据库解决方案 Clustrix

把海量的、无意义的“非结构化数据”进行挖掘提取，整合成结构化数据，并使之有意义或创造价值，这是很多大数据公司的根本愿望。而完成这些任务有一个前提：你必须能从海量数据中找到你需要的那部分，这就是创业公司 Clustrix 正在做的。

Clustrix 在 2010 年曾推出了一个可高度扩容的伸缩式数据库解决方案 Sierra，提供了和 SQL 数据库相似的功能，同时还能对数据存储进行无限制扩展。Clustrix Sierra 被业内称为云计算时代的 MySQL，它可以帮助现在要处理海量数据的公司更快地找到数据并解决日益增长的数据扩容等问题。

Clustrix 创建于 2005 年，在 2006 年曾是 Y Combinator 资助的一个创业项目。2010 年 12 月获红杉资本、USVP、ATA 的 B 轮融资 1200 万美元，2012 年 7 月初又获红杉资本、USVP、ATA 的融资 675 万美元，为接下来的 C 轮融资做准备。Clustrix 公司总部在旧金山，共有 60 人，在西雅图还有研发部门。

Clustrix 网站：http://www.clustrix.com/default.aspx。

大数据高效管理 RainStor

RainStor 由英国国防部的研发小组成立。RainStor 的全球研发中心在美国旧金山，但是核心工程团队仍然在英国，商务开发、市场及直销服务团队在欧洲和美国都有分布以支持合作伙伴和终端用户的客户。

RainStor 是为大企业以最低总成本来管理大数据设计的数据库。RainStor 有两种版本的数据库产品“大数据维护”（Big Data Retention）和“基于 Hadoop

的大数据分析”（Big Data Analytics on Hadoop）来有效管理多个结构化数据集，全面地访问持续查询和分析以帮助满足遵循标准以及最快速地在 Hadoop 上查询和分析。RainStor 创新的数据库相比于传统的关系型数据仓库方式，是以最有效和性价比最高的方式存储和管理多个结构化大数据。RainStor 的专利技术使用复杂数据压缩和重复数据删除技术来减少存储量超过 95%。在 RainStor 上保留的数据能直接用 SQL 进行查询和分析，基于 Hadoop 的 BI 工具或 Map Reduce 不需要再存储或再膨胀的数据。RainStor 以称为区的大块空间存储数据，能使用标准文件系统进行更为方便的管理。HDFS 和低成本存储平台只需很少的资源来新建和保持，可长期减少总成本。RainStor 支持灵活的配置模型，包括云和端，能在包括 NAS、CAS 的广泛的商品硬盘选项中进行配置，也可以使用开源的 Apache Hadoop 分布式文件系统。RainStor 不要求设计、新建索引、调试和持续的维护。

综述，RainStor 能使得组织能以最小的精力、设计存储、硬件和运营节约来存储和分析大规模、虚拟的、没有限制规模的多结构化数据。商业用户能访问这些数据能获得更好的商业洞见。

RainStor 成立于 2004 年，在 2009 年从 Doughty Hanson 和陶氏公司筹集了 400 万美元。那年底公司更名为 RainStor，并在第二轮融资中从 Storm Ventures 和数据集成公司 Informatica 筹集了 750 万美元。2012 年 10 月筹集了 1200 万美元，瑞士信贷和 Rogers Venture Partners 领投。参与融资的还有现有投资者 Doughty Hanson Technology Ventures，Storm Ventures 和陶氏化学公司，截至目前该公司已筹集至少 2350 万美元。

第三节　商 业 智 能

商务数据分析决策 SumAll

很多企业都希望能够借助这样一个日益壮大的平台实现品牌推广，但考虑到社

交网站的数据仍无法找出可行的商业模式，因此这种预期难以全部实现。然而，随着社交数据分析公司的创立和发展，利用社交网站的影响力不再是一个遥不可及的梦想。

纽约的 Sumall 公司，就想要把“小而美”的数据带给每一个人。Sumall 的平台主要是为在线的中小企业提供实时数据服务，可以通过桌面、iPhone 及 Android 访问，它把大量的数据通过可视化形式展现，使得它们直观、易于阅读。同时，在与 Shopify、PayPal 和 Magento（易趣和亚马逊正在使用）电子商务合作伙伴和支付系统合作的 Sumall，用户只需要点击几下即可完成账户集成工作。SumAll 能够快速分析实时数据，然后用一个社交媒体式的“新闻订阅”给用户提供简洁分析和见解。此外，通过 Sumall，客户还能够深入挖掘税收、发货和出售量等数据，甚至是根据不同标准对客户进行排序分析。

Sumall 在 2011 年 11 月成立，2012 年 6 月，由著名风险投资公司 Battery Ventures 牵头，Wellington Partners、Matrix Partners 和 General Catalyst Partners 等跟投为 Sumall 注入 150 万元种子期融资。2012 年 12 月 SumAll 宣布获得 600 万美元的 A 轮融资，此轮投资由 Battery Ventures 领投，Wellington Patners 参与投资。公司目前拥有 25 名员工，全部在纽约总部。

Sumall 网站：https://sumall.com/。

敏捷数据管理 Delphix

大数据和云计算，对于消费者来说是个无聊的东西，但是对于利用这两种热门的后端技术进行敏捷数据管理的 Delphix 来说却是一座大金矿。

Delphix 的敏捷数据管理解决方案不需要部署冗余的基础设施，同时可以加快相关流程。这样客户就可以更快更省钱地交付应用。所谓敏捷数据管理就是指在企业数据库内对数据进行虚拟化，从而提高数据库驱动型应用的开发的敏捷性，这会令数据库及应用管理的面目焕然一新。Delphix 会把企业的数据库放到云上面，利

用数据同步和虚拟化技术将合适的数据交到恰当的人手里。Delphix 声称采用其应用交付解决方案的应用项目进度可提高 5 倍，成本可节约 90%，这家公司自 2010 年面世以来的销售年增长率为 300%。

Delphix 成立于 2010 年，在 2012 年 6 月完成了其 C 轮融资 2500 万美元。此轮融资由 Jafco Ventures 领投，Greylock Partners 等亦有参与。目前 Delphix 的总融资额已达 4550 万美元。该公司主要依靠其“敏捷数据”获得了超额认购。“敏捷数据”通过虚拟化企业数据库的数据，其增加了数据驱动应用的敏捷性；提高了经济数据库和应用管理速度。

Delphix 网站：http://www.delphix.com/。

数据智能解决方案 Ngdata

Ngdata 能让企业用户和他们的消费者通过先进的一对一营销方式，作出更好的建议和产品。Ngdata 的产品 Lily 将把企业的内外部结构化／非结构化数据集成在一个平台上。Lily 使用人工智能拍照工具记录消费者们的习惯和喜好。大数据市场正在快速增长，它对企业的意义越来越重要，企业可以通过它提供的分析数据快速评估市场和采取行动。ING 投资总监 Tom Bousmans 称，消费者每秒产生上亿数据，这些数据能让企业有机会更好地了解用户需求，与用户开展个性化的、动态的互动。

Ngdata 成立于 2009 年，目前有 20 名员工。竞争对手包括 Wibidata 和 Spire。不过与竞争对手不同的是，Ngdata 称其提供的数据解决方案使得企业可以与消费者实现互动，而非单纯专注于大批量的数据分析。NGDATA 于 2012 年 10 月融资 250 万美元。Lily 本次融资资金来自于 ING、SniperInvestment、Plug and Play Ventures 等投资机构和一些天使投资人，资金将有助于 Ngdata 拓展个性化产品线和为美国客户设立纽约及旧金山服务办公室。

Ngdata 网站：http://www.lilyproject.org/lily/index.html。

智能搜索引擎 Attivio

Attivio 创始人 Ali Riaz 认为当企业用户发送一条查询请求时，得到的应是有洞察性的信息，而不是罗列链接或是仅给出一张图表。要解答出“为什么”而不仅是“是什么”，比如要能分析出销售下降是因为市场需求下降还是因为销售人员表现不够突出等。

Attivio 的核心产品是 AIE（Active Intelligence Engine），一个智能引擎。这个 AIE 将企业的结构化和非结构化的各类数据整合起来，形成统一的信息接入平台，让企业人员可以方便地检索和分析信息。这就弥补了原先企业商业智能只分析结构化数据而忽略非结构化数据（如邮件及各类商业文档）中大量有价值信息的缺憾。

Attivio 成立于 2007 年，创始人及 CEO 为 Riaz，他曾是 FAST Search & Transfer 的总裁。FAST 这家公司 2008 年被微软以 12 亿美元的价格收购。Attivio 在 2012 年 10 月获 A 轮融资 3400 万美元，这轮融资由 Oak Investment Partners 领投。

Attivio 网站：http://www.attivio.com。

数据分析决策集成服务 GoodData

GoodData 提供的是基于云的数据分析服务，但其竞争对手都是一些业界巨头，包括 IBM、SAP 和 Oracle 等。不过，GoodData 的优势是商业模式。跟那些巨头提供的套件式解决方案不同的是，GoodData 向广大的 SaaS 提供商提供技术集成服务，让他们在自己的平台中集成其数据分析技术，从而使得这些 SaaS 提供商可以向最终客户提供诸如仪表盘、报表等功能。

市场营销是任何企业都需要做的工作。最近几年，由于社会化媒体的兴起，数字营销逐步成为营销业者关注的焦点，但是营销人员对这个领域仍缺乏有效的分析。因此 GoodData 瞄准了这一点，利用集成服务为营销人员提供对微博、社交网络及在线营销活动的深度分析功能。

GoodData 公司创立于 2009 年，2011 年 8 月 B 轮融资 1500 万美元，Andreessen Horowitz 领投。2012 年 7 月 GoodData 又融资 2500 万美元，投资者包括 Andreessen Horowitz、General Catalyst Partners 及 Fidelity Growth Partners，总融资达 5500 万美元。

GoodData 网站：http://www.gooddata.com/。

第四节 垂 直 应 用

云笔记存储服务 Evernote

Evernote 是一款非常便捷的应用，以文字、图片、语音、网页文本截取等多种格式帮用户记录身边的一切，并可将多种格式转换为文字，进而可在不同的记录格式中快速搜索。所有的信息都会同步到 Evernote 云服务器，可在 Windows、Web、iPhone、iPad、Android、黑莓、Windows Mobile 等多个平台上跨平台使用。简单来说，Evernote 就像一个功能强大的魔法笔记本，用它可以做日程便签、笔记摘要、资料采集器，甚至可以收藏和制作富媒体文件。用 Evernote 公司自己的话说就是：Remember Everything—Capture anything，Access anything，Find things fast（记住一切——捕捉一切，便捷接入，快速查找）。

Evernote 于 2008 年成立，发展的注册用户数随时间变化见表 13-2。

表 13-2 Evernote 发展的注册用户数随时间的变化情况

时 间	注册用户数
2009 年 5 月	100 万
2009 年 12 月	200 万
2010 年 5 月	300 万
2010 年 8 月	400 万
2010 年 11 月	500 万
2011 年 6 月	1000 万
2011 年年底	2000 万

2012 年 5 月，云笔记存储服务 Evernote 完成了 D 轮融资，此次融资金额为 7000 万美元，此轮融资由 Meritech Capital 和 CBC Capital 牵头，对 Evernote 的估值也达到了 10 亿美元。2012 年 12 月，Evernote 完成新一轮 8500 万美元的融资。其中 75% 为现有投资者二次注资，25% 为新投资者。AGC Equity Partners/m8 Capital 主投，T. Rowe Price 跟投。

Evernote 网站：http://evernote.com/intl/zh-cn/。

内部销售和线索反馈管理 InsideSales

作为内部销售和线索反馈管理行业的领导者，InsideSales 为想要转换生产及销售团队职业性的机构提供了软件、培训及咨询解决方案。InsideSales 研究报告被发表在哈佛商业评论上。

InsideSales 成立于 2005 年，现在美国犹他州的普罗沃。InsideSales.com 在 2012 年 8 月完成了 A 轮融资，从 Hummer Winblad 等公司筹集了 400 万美元。据悉此次 InsideSales.com 获得的融资主要用来拓展数据解析及销售自动化技术。参与这轮融资的还有 Omniture 联合创始人和前首席执行官约什 • 詹姆斯 （Josh James）。在这次融资之前，该公司是盈利的，并没有获得任何投资。InsideSales 的竞争对手包括了 Leads360 和 Five9。

InsideSales 网站：http://www.insidesales.com/。

医疗保健行业大数据解决方案 Predilytics

Predilytics 是一家提供医疗行业解决方案的信息技术公司。公司的产品聚焦于协助健康计划，健康提供及风险规避的识别及优化机会，优化疾病负担的合适的文件，吸引和保留计划，提高关怀管理的有效性，减少高昂贵的成本设施。

Predilytics 公司的产品协助用户进行健康计划，让医疗保健行业的供应商和承

担风险的团队识别和优化核心商业机会，吸引和保留疾病患者存档，同时提高医疗保健管理的效率，降低高额设备的授权和重新授权费用。Predilytics 利用了在金融和广告行业广受认可的最新的机器学习技术，而这些技术至今还没有在医疗保健市场大规模使用。Predilytics 让包括医院和保险商在内的从业者提高运营效率，提高收入，从而为行业的发展打下基础。

Predilytics 在 2012 年 8 月已进行 A 轮融资共 600 万美元，投资方为 Flybridge Capital，Highland Capital 和 Google Ventures。所筹资金将会被用于未来产品扩张和运营发展。

Predilytics 网站：http://www.predilytics.com/。

职业搜索引擎 TalentBin

TalentBin 是一个信息综合分析网站，通过收集人们日常在社交网络上的信息，建立一个以人为中心的数据库。日前，它在 iPhone 平台上发布了一款新的以人为中心的搜索引擎，主要的搜索内容和网站的基本类似。TalentBin 的人才搜寻服务器，通过利用人们浏览网络所留下的信息来创造“隐形的简历”(implicit resumes)。将其整合，放入一个一站式的搜索平台，使得猎头可以找到并发掘这些申请人。

TalentBin 成立于 2011 年 5 月，总部设在美国旧金山，拥有员工 15 名。2012 年 5 月，TalentBin (其前身为 Honestly.com) 就为企业招聘人员推出了一款人才搜索引擎。2012 年 9 月，TalentBin 推出了该搜索引擎的 iPhone 版本，但它的受众并非仅仅针对招聘者，也包含那些想要寻找他人的用户。2012 年 9 月 TalentBin 获融资 1000 万美元，利用新融资，TalentBin 已开始探索除社交网络活动以外的其他领域。

TalentBin 网站：http://www.talentbin.com/。

第五节　其　　他

大数据分析应用 Datameer

Datameer由一些Apache Hadoop的原始创立者创建。Hadoop是由Apache基金会开发一个分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用机群的威力高速运算和存储。Hadoop 分布式文件系统（Hadoop Distributed File System，简称HDFS）有着高容错性的特点，并且设计用来部署在低廉的硬件上。HDFS放宽了可移植操作系统接口（POSIX）的要求，能以流的形式访问文件系统中的数据。它提供高传输率（high throughput）来访问应用程序的数据，适合那些有着超大数据集的应用程序。

现在Datameer已经成长为一个全球团队，专注于高级的大数据分析。在几次“世界500强公司内的Hadoop分析解决方案”完成后，创始人决定建建下一代分析应用来解决在结构化和非结构化开发过程中所创造的新的用户案例。Datameer是通过联合数据集成、数据转换、数据可视化进行大数据分析的单一应用。传统的商业智能系统要求一个复杂的、多步骤的、多渠道的过程并且受限于特定用户的反复训练集。Datameer提供了一个简单的大数据分析应用，不要求ETL（数据提取、转换和加载），无静态模式，并能将有用的分析和数据可视化展示给任何用户手中。以前客户需要ETL（数据提取、转换和加载）、静态数据仓库和IT驱动的商业智能三个步骤过程的数据分析，分析过程包括三个不同渠道三组专家团队和三种不同的技术。Datameer将这一复杂的环境简化到强大的Hadoop平台上的简单应用，可轻松实现数据集成、分析和可视化。

Datameer成立于2009年，在2011年5月融资925万美元，由KPCB领投，Redpoint Ventures跟投。总投资已达1200万美元。

Datameer 网站：http://www.datameer.com/company/index.html。

跨平台大数据处理 Trifacta

Trifacta 认为挖掘数据的价值并不是从数据或计算机获取的，而是来源于人——分析师或决策者能抽取出洞见和优化过程。“价值来源于人”是简单的经济学。随着存储和计算成本的下降，技巧分析的成本的上升，以及数据的多样性和规模的增加，人的专业性需求在急速增长。今天经济学稀缺的资源很明显：能掌握数据和计算的人。这些经济事实定义了巨大的挑战：迅速掌握数据分析的生产力。需要通过跨人机交互界面、规模化数据管理和机器学习等跨领域的洞见，要求人、数据和计算在内的解决问题的新技术来满足这一挑战。

Trifacta 成立是为了解决面临的分析数据能力的挑战。基于伯克利和斯坦福多年的合　作研究，Trifacta 在关键领域联合了技术领先者。Trifacta 团队在重新思考帮助用户控制数据的用户界面、系统和算法。这是一个令人兴奋的组合，也是一个简单的目标：在数据分析领域建立直观、强大和有用的解决方案。

Trifacta 跟其他一些致力于简化大数据使用的公司不同，其关注点是创建可供多个不同平台（传统的关系式数据库、Hadoop 集群）使用的接口。Trifacta 的作用是创建可在多个实体数据存储及处理系统上运行的 SQL 查询或 map reduce 代码。

Trifacta 的目标客户不仅包括普通的商业用户，数据科学家也是其争取的对象。用户可取出数据集或其样本到内存中，然后用 Trifacta 通过接口利用多种可视化数据方式来浏览这些数据。应用还可以对客户下一步的操作提出若干建议，操作执行的效果还可以预览。一旦决定了希望执行的操作，相应的代码或查询就可以生成。可视化可以帮助突破大数据技术中人遇到的瓶颈，软硬件再加上人的力量，大数据爆发指日可待。

Trifacta 在 2012 年 10 月获得 Accel Partners Big Data Fund 的 430 万美元投资。

Trifacta 网站：http://trifacta.com/。

数据分析平台 Precog

Precog 是一个帮助开发者和数据科学家抓取、丰富和深入分析大量多种结构数据的平台，能在他们的应用中获取强大的洞见和智慧。Precog 更快和更方便地将数据资产植入数据产品中。Precog 是一个云平台，不需要客户安装或维护，无论数据或查询的多少，服务器都可扩大规模以满足客户需求。

Precog 提供了一个数据删除、增加和处理的集成市场。在控制面板中少量点击，客户即可使用任何存储在 Precog 上的数据，以及使用情绪分析、人口统计学分析等功能来删除、增加和处理数据。Precog 提供了固定模式的分析和统计，当客户仅仅基于如 Hadoop 或 MongoDB 的数据存储创建功能时，它会花费很多的时间和很大的精力，Precog 使得客户可以不需要太多精力而创建大量的功能，从而使客户更容易聚焦于问题，而不是技术上。

Precog 的成立于 2010 年，于 2012 年 2 月由 CEO John A De Goes 宣布正式成立，注册资本 277 万美金。2012 年 5 月获得 RTP 风险投资的 200 万美元，同时 RTP 高级经理 Kirill Sheynkman 进入 Precog 公司董事会。

Precog 网站：http://www.precog.com/。

实时大数据分析 ParStream

在 ParStream 创始人发现客户在实时大数据应用时缺乏数据库技术，他们将超过 15 年的 IT 行业职业经历与埃森哲的职业经验相结合，从现有服务业务 empulse 出发，启动了 ParStream 的开发，一个值得信赖的产品公司由此诞生。

ParStream 的目标是要在数据库市场实现实时大数据分析应用，ParStream

为数据库数据技术进行了基础研究和驱动创新，允许用户以明显的低成本进行大数据的实时数据分析。ParStream 压缩了数据，从而转化为对小规模数据进行实时分析。由于该公司与惠普的 Vertica 和 EMC 的 Greenplum 竞争，可能成为收购目标，Vertica 和 Greenplum 都曾是独立的公司。

ParStream 成立于 2008 年，原先位于德国，现在在科隆和帕罗奥尔托都设立了分支机构。2012 年 8 月，ParStream 在首轮融资中筹集了 560 万美元，以进一步开发其使用内存和大型并行处理进行实时分析的大数据分析技术。这次融资由 Khosla Ventures 领头，参与融资的还有 Baker Capital、Crunch Fund、Data Collective、Tola Capital 和一些个人投资者。

ParStream 网站：http://www.parstream.com/en/home/index.html。

用户行为监测分析 Mixpanel

Mixpanel 建设有 Web 和移动分析平台，其提供的服务可以分析监测用户活动。从 2009 年 7 月拿到种子资金至今，他们一直保持着高速的发展：2010 年 10 月 Mixpanel 发布邮件分析工具，2010 年 11 月允许开发者向用户提供即时分析数据，2011 年 1 月 Mixpanel UI 大转变，月数据量增长 40%。2011 年 6 月，新产品 Mixpanel Streams 实现了实时监测用户在客户网站上的活动。

Mixpanel 成立于 2009 年，投资者清单令人炫目。此前已从红杉资本、Square COO Keith Rabois、PayPal 联合创始人 Max Levchin 等处获得 175 万美元的融资。2012 年 5 月，Mixpanel 获 Andreessen Horowitz 领投的 1025 万美元 A 轮融资。这次跟 Andreessen Horowitz 一起参与 A 轮融资的还包括了 Salesforce.com CEOMarc Benioff 以及 Yammer CEO David Sacks。

Mixpanel 网站：https://mixpanel.com/。

数据库服务商 DataStax

DataStax 支持大数据 apps，大数据 apps 已为超过 200 家客户转换业务，包

括初创公司及财富 100 强中的 20 家公司。DataStax 在 Apache Cassandra 上建立了大规模、富有弹性和可持续使用的大数据平台。DataStax 在云端跨多数据中心为分析集成了企业级 Cassandra、Apache Hadoop，为研究集成了 Apache Solr。

DataStax 的产品聚焦于使客户轻松地创建和运营由 Apache Cassandra 支持的大数据设备，帮助客户在复杂的大数据世界里可以更迅速行动。通过 DataStax，客户可以集中于大数据应用，而不用考虑基础架构的复杂性或不足。Apache Cassandra 是大规模开源的关 NoSQL 数据库，是设计用于通过跨多个数据中心和云来持续提供的掌握大规模数据的 Apache 软件基础水平的项目。Cassandra 从谷歌、亚马逊和 Facebook 的工作进化而来，并被领先的公司如 Netflix、Rackspace 和 eBay 使用。

DataStax 创立于 2010 年，是大数据时代下诞生的创业公司，2011 年完成 1100 万美元 B 轮融资，本次融资由 Crosslink Capital 和 Lightspeed Venture Partners 领投；2012 年，Data Stax 完成了 C 轮 2500 万美元的融资。

DataStax 网站：www.datastax.com。

附录　大数据发展大事记

时间	大数据事件	里程碑
2011 年 5 月	麦肯锡全球研究报告《Big data: The next frontier for innovation, competition,and productivity》	首开先河
2011 年 5 月	EMC World 2011 在拉斯维加斯开幕，会议主题为“云计算适逢大数据”，参会者超过 10000 人,现场有超过 500 场讲座，以及来自上百家领先 IT 厂商的上百个动手实验室和展示。EMC 公司董事长兼首席执行官乔图斯先生发表主题演讲为四天的大会开幕，他着重介绍了云计算和大数据给 IT 带来的变革。同期举办 Momentum 大会（企业内容管理大会）、数据科学家峰会（Data Scientist Summit）、大数据存储峰会和 CIO 峰会	
2011 年 5 月	IBM 推出大数据分析软件平台 InfoSphere BigInsights 和 Streams，这是目前业内最先推出的针对大数据分析的产品。两款产品将包括 Hadoop MapReduce 在内的开源技术紧密地与 IBM 系统集成起来。研究 Hadoop 这样开源技术的人很多，但是 IBM 这次是真正将其变成了企业级的应用	
2011 年 7 月	Yahoo 宣布成立新公司 Hortonworks 接手 Hadoop 服务，Hadoop 也迎来了新的发展机会。针对大数据领域，其实有很多技术提供商都参与了 Yahoo 的项目。Apache Hadoop 是一个开源项目，Yahoo 就是其中最大的贡献者;Google MapReduce 是 Hadoop 架构的一个主要组件，开发出的软件可以用来分析大数据集，它在目前的火爆程度已经无需赘言;Cloudera 是 Hadoop 最早的技术支持、服务和软件提供商，它今后将直接与 Yahoo 的 Hortonworks 展开竞争。此外，EMC 还推出了付费的 Hadoop 产品并基于 MapR Technologies 公司的技术	

续表

时间	大数据事件	里程碑
2011 年 8 月	微软宣布推出了两个基于 Hadoop 的大数据处理的社区技术预览版连接器组件，一个用于 SQL Server，另一个用于 SQL Server 并行数据仓库（PDW）。该连接器是一个部署在 Linux 环境中的命令行工具。SQL Server Hadoop 连接器在微软大数据之路上最重要的一步。另外，微软还宣布将推出 LINQ Pack、LINQ to HPC、Project“Daytona”以及 Excel DataScope，这些产品都将专为研究人员和业务分析师打造，用以在 Windows Azure 上做大数据分析	
2011 年 10 月	甲骨文宣布收购为企业用户提供非结构化数据管理、网络商务和商务智能技术的企业搜索和数据管理公司 Endeca Technologies	
2011 年 12 月	中国资本市场发布第一篇大数据主题研究报告《大数据时代即将到来》	掀起资本市场热潮
2012 年 1 月	中国资本市场发布第二篇大数据主题研究报告《大数据时代三大发展趋势和投资方向》	提出系统的大数据认知框架
2012 年 3 月	IDC 发布大数据（Big Data）市场预测报告，预估该领域的市场规模将从 2010 年的 32 亿美元成长到 2015 年的 169 亿美元，每年的平均成长率接近 40%	
2012 年 3 月	亚马逊 CTO Werner Vogels 在 Cebit 上发表的主题演讲“无限的数据”时称，企业在思考大数据的时候，需要注意的不仅是需要分析大量的数据，还包括信息的存储方式。此外，还鼓励企业思考大容量图片的问题，他还介绍了用于实施大数据系统的亚马逊云蓝图	
2012 年 3 月	美国奥巴马政府宣布推出“大数据的研究和发展计划”。该计划涉及美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美国国防部高级研究计划局、美国地质勘探局等 6 个联邦政府部门，承诺将投资两亿多美元，大力推动和改善与大数据相关的收集、组织和分析工具及技术，以推进从大量的、复杂的数据集合中获取知识和洞见的能力。美国奥巴马政府宣布投资大数据领域，是大数据从商业行为上升到国家战略的分水岭，表明大数据被正式提升到战略层面，大数据在经济社会各个层面、各个领域都开始受到重视	标志大数据上升为国家战略，体现国家意志

续表

时间	大数据事件	里程碑
2012 年 4 月	中国资本市场发布第三篇大数据主题研究报告《以数据资产为核心的商业模式》	系统阐述大数据商业图景
2012 年 4 月	SAP 计划斥资近 5 亿美元来吸引用户使用其 Hana 数据处理产品，从而加大与甲骨文之间的竞争。Hana 平台的设计目的是迅速分析海量的销售和运营信息，以及对电子邮件和社交媒体等非结构化数据进行分析，依靠计算机存储器而非磁盘驱动器来加速这一程序。	
2012 年 4 月	企业数据软件公司 Splunk 今天以每股 17 美元的价格在纳斯达克进行 IPO，融资 2.3 亿美元，首个交易日市值突破惊人的 30 亿美元	首家上市的大数据公司
2012 年 4 月	谷歌正式推出在线存储服务 Google Drive	
2012 年 5 月	Google 推出的一项企业级大数据分析的云服务 BigQuery，用来在云端处理大数据。BigQuery 将有助于企业在没有硬件基础设施的情况下分析他们的数据。同时可以建立应用程序和数据共享的所有服务	
2012 年 5 月	IDC 发布研究报告指出，“大数据”概念正在引领中国互联网行业新一轮的技术浪潮，截至 2011 年底，中国互联网行业持有的数据总量已达到 1.9EB(1EB 艾字节相当于 10 亿 GB ）。IDC 预计，这一规模到 2015 年将增长到 8.2EB 以上	
2012 年 6 月	IDC 发布研究报告《中国互联网市场洞见：互联网大数据技术创新研究，2012》，对中国互联网行业围绕大数据的技术创新进行了专题研究。报告指出，大数据正在引领中国互联网行业新一轮的技术浪潮	
2012 年 6 月	为推动大数据（Big Data）这个交叉学科的发展，推动学术、应用和产业的发展，中国计算机学会决定成立“CCF 大数据专家委员会”（暂定），并责成 CCF 名誉理事长、中国工程院院士李国杰教授作为牵头人，开展有关工作	
2012 年 7 月	联合国在纽约发布了一份关于大数据政务的白皮书《大数据促发展：挑战与机遇》，总结了各国政府如何利用大数据更好地服务和保护人民	

续表

时间	大数据事件	里程碑
2012 年 9 月	北京拓尔思信息技术股份有限公司联手华为技术有限公司倾力推出拓尔思-华为大数据一体机系列。拓尔思-华为大数据一体机系列包括拓尔思-华为信息采集一体机、拓尔思-华为检索一体机，后续还会有相应的大数据一体机问世	
2012 年 10 月	中国通信学会大数据专家委员会成立大会暨首届大数据论坛在北京召开。会上，大数据的飞速发展以及给我国相关产业带来的机遇和挑战成为与会专家讨论的焦点	
2012 年 10 月	市场研究公司 Gartner 发布研究报告称，大数据产业今年将在全球范围内带来近千亿美元的 IT 开支。Gartner 在报告中预测，今年，大数据对全球 IT 开支的直接或间接推动将达 960 亿美元；到 2016 年，这一数字预计将达到 2320 亿美元	
2012 年 10 月	IBM 和牛津大学联合发布了一份大数据研究报告。研究包括：大数据的实际使用情况；创新型企业如何从不确定数据中提取有价值数据	
2012 年 11 月	淘宝和天猫今年的交易总额在 11 月 30 日突破 1 万亿元人民币，为支撑这巨大规模业务量的直接间接就业人员，已经超过 1000 万	
2012 年 11 月	首届数据科学与信息产业大会在北京国际数学研究中心召开。标志学术界、产业界、资本市场形成共识	数据科学登上产业舞台

后记 POSTSCRIPT

虽然和出版社早早签了合同，但却迟迟没有动笔。第一，杂事较多，静不下心。资本圈的朋友肯定都了解，我们做卖方分析师的，每天就是跑来跑去地调研、路演、电话。一头牵着上市公司，一头牵着投资机构，每日奔忙中又看着股价上蹿下跳，未免跟着一喜一忧。不知有止，无定无静。就算是动笔，也是虚浮跳脱的文字。第二，这是副业，本想利用业余的时间方好，所以签约的时候，就推到年底交稿。计划在股票市场淡季的时候突击。

不料写作难度，超过了最初的想象。作几页供演讲的幻灯片、编写一份报告和写作一部书的难度，确乎不可同日而语。原本打算就用几篇报告打个底子，找些兄弟搜罗点资料，然后再拼凑一些公司介绍也就能交差了。不想动笔之后，却发现最初的想法太幼稚。

写作是再学习的过程

“大数据”虽然 2012 年迅速“蹿红”，但是却莫衷一是，千人千面。媒体报道虽然热热闹闹，深度资料却又少之又少。于是，每一章的写作，都是对某个行业、某个陌生主题、某个陌生领域的苦苦求索。就像在崎岖的地方走夜路，仅仅拿一盏

纸糊的灯笼照亮，刚刚看清脚下的路，猛抬头，又陷入无边的黑暗之中。

本书的阅读对象，锁定为各行各业的企业高管，资产市场上的新兴趋势感兴趣的投资人。所以，必须要勾勒出一个行业的发展脉络、驱动要素、竞争制高点。如此才能深入浅出地洞察大数据在其中的作用。

写作过程中遇到的困难林林总总。譬如媒体，需要系统地了解传媒、广告、营销等领域的知识。而且我也不打算用过多的专业术语。因此自己归纳总结了发展脉络。关于媒体的这章写完，我不确定是否读者能接受文中观点，但是我对传媒的理解，倒是上升了一个台阶。

再如第三章提到的数据资产，尽管一年多来，人人都认为数据资产重要，有人把他比作石油，有人将它比作血液。但从来没有人说符合哪些特征的数据资产是比较好的？我们查遍海内外资料，阅读新出炉的专著，也没有找到这方面的只言片语。如果这些特征不能提炼出来，那么我们面临投资抉择的时候，就无法去对比衡量一家公司的数据资产，无法评估数据资产的潜在价值。因此，只能不断去探索、发现和总结提炼。当写完这本书，后头再审视当时提出的“数据资产评估模型”，发觉也不过尔尔。因为这个模型只是定性的给出五个维度，仅仅是考察数据资产的思维框架。如果要付诸实用，还需要不断地补充、完善。

边写、边学、边思考。这本书，就是这段时间思考的结果。

写作也是广交朋友的过程

写作不是目的，在写作中遇到志同道合的朋友，才是一大乐事。在这个意义上，这本书作为一个载体，让我们结识了更多的人，给了我们更宽广的视野。在一次技术沙龙上，碰到糜博士。不巧糜博士还在感冒，尽管声音低沉，却掩饰不了他对大数据的深刻洞见。惺惺相惜之际，他也成为这本书的合作者之一。

因为职业的原因，上市公司高管是我们很重要的伙伴。但是通常而言，分析师一般只和董秘打交道，和公司的 CTO、CIO 们缺少共同的语言。但是大数据是我们和企业技术负责人之间非常自然的一个话题。彼此启发，相互印证，乐趣无穷。

前段时间参加某银行内部的质量保障年会，我受邀给大家讲讲时下热点。没想到大数据非常受欢迎。原来光大银行早就成立“数据治理中心”来统筹全行的数据发展战略，一下子和这个中心的主任找到了共同兴趣点。这家银行就此成为我们观察行业发展趋势的窗口。

鄂维南院士的支持，为我们注入强大的动力。鄂老师一直致力于推动应用数学和产业界深入合作。数据科学就是在学术界和产业界深度融合的背景下提出的。鄂老师的思想，和我们不谋而合，于是欣然动笔写作数据科学部分。鄂老师治学非常严谨，把书稿全篇打印出来，统揽全书之余，又逐字逐句地校对。单这种精神，就值得学习。

网友的支持是促使我决心写作的一个重要原因。2012 年初，网友@尹锴_ink看到我的一篇博客文章《大数据时代的三大发展趋势和投资方向》中提到六种商业模式。热心指出模式虽然罗列的挺全面，但是不同的模式之间缺乏结构性的联系。也就是罗列有余，分类不足。尹锴亲自动手，给我做了 7 幅图。这些图都被用在我的报告中，其中第一幅用在本书第五章。尹锴是甲骨文北美的资深顾问，一直在美国生活，到现在还没有谋面。我们一直靠微博联系。夏明武是促使我在微博上公开写书进展的关键角色。把他在工作中沉淀的大数据思想和实践，毫无保留地告诉我。让我对电信运营商数据资产的价值，有了全新的认识。还有许多朋友，有的提供案例线索，有的提供公司介绍，或者仅仅简单的两个字“期待”，就足以让我加班加点地写作。

打字的速度跟不上大数据发展的脚步

按照构想，这本书只是勾勒出大数据的认知框架。以数据资产为核心，遍历各行各业，洞悉发展趋势，发掘投资标的。所以如果按照不同的行业来写，真是不知道什么时候能够写完。因为各行各业都开始谈论这件事情，有走得快的，都已经任命了专门的高管。大多数都处在好奇、探究的阶段。所以书中只是涉及了那些走在前列的行业，譬如媒体、再如金融。

事实上，累积数据资产并善加运营，是各个行业的事情，甚至是未来五年到十

年重头戏。所以，我们还有很多的行业没有涉及。智慧城市、电子政务、医疗保健等都是大数据的金矿。安全是一个大主题，大数据时代，数据的重要性，上升到资产层面。面对海量的数据资产，安全领域该如何演进，我们也没有展开。只是在概述中地略地提到保护个人隐私和加强数据安全的重要性。限于时间和精力，第一版没有包含这些领域的探讨。如果得到读者的好评，我们计划在第二版中补充一些行业的案例。

大数据领域发展变化之迅速，的确令人眼花缭乱。刚刚完成书稿，就看到两则新闻。其一是苹果的 AppStore 下载量超过 400 亿次，仅仅 2012 一年，就下载了 200 亿次。另外一则关于投资领域的简单统计数据：根据资本实验室 2012 年度创业投资与并购报告，2012 年国内外数据服务领域以 124 起投资案例在所有行业中排名第二。这一数据表明，“大数据”不再只是空泛的概念，而是以非常快的速度在各领域落地生根。在 100 多起案例中，除了大数据与社交媒体、电子商务、广告/营销等行业的紧密融合，特别值得关注的是专业化的数据服务已经渗透到农业、建筑、能源、体育、餐饮、音乐等传统行业，而这一趋势在未来将会得到进一步强化，并将极大推动传统产业的升级；对于新兴数据服务商来说，也是非常值得关注和把握的商业机会。

按照这个趋势发展下去，我在电脑旁打字的速度，无法追上大数据发展的脚步。临近年关，践约交稿，欲说还休的其他话题，留待下一版分解吧！

赵国栋

2013 年 1 月于北京

参考文献

[1]《大数据研究和发展计划》,(网络地址:http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal).

[2] 国金证券大数据系列研究报告第一篇《大数据时代即将到来》.

[3] 国金证券大数据系列研究报告第二篇《大数据时代的三大发展趋势和投资方向》.

[4] 国金证券大数据系列研究报告第三篇《以数据资产为核心的商业模式》.

[5] 美国国防部立项的几个大数据项目(原文参见网络地址:http://www.white-house. gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf).

[6] Plattner，Zeier. In-Memory Data Management, 2011.

[7] Driscoll.Big Data Now.

[8] 谢耘.转折——IT 产业透视.

[9] Katy Huberty, Ehud Gelblum. Morgan Stanley Research.Data and Estimates as of 9/12.

[10] 麦肯锡.Big data: The next frontier for innovation, competition, and productivity,2011.

[11] www.chrisabraham.com.

[12] 互联网数据中心的《中国互联网市场洞见：互联网大数据技术创新研究2012》.

[13] 陆迪的博客（网络地址：http://blog.sina.com.cn/s/blog_53f9871e0101ewbh.html）.

[14] THE STATE OF THE INTERNET（网络地址：http://www.businessinsider.com/state-of-internet-slides-2012-10?op=1）.

[15] MAPPING THE DISPLAY LANDSCAPE（网络地址：www.netmining.com/displayguide）.

[16] 吴军. 浪潮之巅[M]. 北京：电子工业出版社，2011.

[17] 美国广告产业链（网络地址：http://news.iresearch.cn/Zt/153133.shtml#a2）.

[18] 解析 DSP（需求方平台）的意义（网络地址：http://ljq19841984.blog.163.com/blog/ static/133020519201211095826629）.

[19] Google Remarketing 再营销详解_百度文库.

[20] 中国互联网广告领域 2012 变化总结（网络地址：http://a.iresearch.cn/bm/20121205/188375.shtml）.

[21] 用户定向数据的现况和未来（引自 BCG 咨询公司）| The Evolution of Online-User Data（网络地址：http://www.rtbchina.com/the-evolution-of-online-user-data.html）.

[22] 互联网精准广告定向技术（网络地址：http://www.iamniu.com/2012/05/26/summary-internet-precise-ad-targeting-technology）.

[23] Pixazza 把每张图片自动变成广告赚钱（网络地址：http://www.alibuybuy.com/posts/4351.html）.

[24] 视频广告新思路-点击画面即可购买（网络地址：http://www.alibuybuy.

com/ posts/27355.html）.

[25] Enprecis 解决方案.

[26] 谷歌公司的收购史维基百科.

[27] eMarketer digital Intelligence，Digital Ad Trends，2012.

[28] IAB,U.S.Census Bureau,Stategy Analytics.

[29] 谷歌公司 2012 年第三季度数据整理资料（网络地址：http://www.wordstream.com/blog/ws/2012/10/25/google-faces）.

[30] eMarketer digital intelligence 2012 “digital ad trends”.

[31] 推荐引擎初探（网络地址：https://www.ibm.com/developerworks/cn/web/1103_ zhaoct_recommstudy1/）.

[32] 布劳尔 • 约翰 • F • 肯尼迪.

[33] 德鲁克 • 管理的实践[M].北京：机械工业出版社，2009.

[34] 亚德里安 • 斯莱沃斯基等.发现利润区[M]. 北京：中信出版社，2010.

[35] 亚历山大 • 奥斯特瓦德等.商业模式新生代[M]. 北京：机械工业出版社，2012.

[36] 《麦肯锡季刊》中文版（china.mckinseyquarterly.com）.

[37] 黄家明，方卫东.交易费用理论:从科斯到威廉姆森[J].合肥工业大学学报（社会科学版），2000.

[38] http://www.data.gov.

[39] https://github.com/opengovplatform/opengovplatform.

[40] 丁健.浅析大数据对政府 2.0 的推进作用[J]中国信息界，2012.

[41] http://open-data.europa.eu/open-data.

[42] 大数据产业全景图 http://blogs.forbes.com/davefeinleib.

[43] DataSift 网站：http://datasift.com/.

[44] Evernote 网站：http://evernote.com/intl/zh-cn/.

[45] Junar 网站：http://www.junar.com/.

[46] Precog 网站：http://www.precog.com/.

[47] 10gen 网站：http://www.10gen.com/.

[48] Nirvanix 网站：www.nirvanix.com.

[49] Mixpanel 网站：https://mixpanel.com/.

[50] Sumall 网站：https://sumall.com/.

[51] Delphix 网站：http://www.delphix.com/.

[52] Clustrix 网站：http://www.clustrix.com/default.aspx.

[53] GoodData 网站：http://www.gooddata.com/.

[54] ParStream 网站：http://www.parstream.com/en/home/index.html.

[55] InsideSales 网站：http://www.insidesales.com/ .

[56] TalentBin 网站：http://www.talentbin.com/.

[57] Predilytics 网站：http://www.predilytics.com/.

[58] Datameer 网站：http://www.datameer.com/company/index.html.

[59] Trifacta 网站：http://trifacta.com/.

[60] Ngdata 网站：http://www.lilyproject.org/lily/index.html.

[61] RainStor 官方网站：http://rainstor.com/.

[62] DataStax 网站：www.datastax.com/.

[63] Attivio 网站：http://www.attivio.com/.

[64] 2011 年数据库市场盘点之大数据，http://www.searchdatabase.com.cn/showcontent_56803.htm.

[65] 国金证券计算机行业日报.

[66] 国金证券计算机软件行业研究月报.

[67] Jun Z.Li，et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation[J].Science,22,Feburary，2008.

[68] Julie Bort. Facebook Stores 240 Billion Photos And Adds 350 million More A Day. Jan. 2013. http://www.businessinsider.com/facebook-stores-240-billion-photos-2013-1.

[69] 陈嘉恒，Hadoop 实战[M]. 北京：机械工业出版社，2011.

[70] 董思颖.Facebook 开发的 HDFS 和 HBase 新特性.Hadoop 与大数据技术大会，2012，11.

[71] 大数据案例分析:电信业 Hadoop 应用分析,http://datacenter.watchstor.com/news138619.htm.

[72] Cashcow.企业需要什么样的数据科学家，http://www.ctocio.com/management/career/5394.html.

[73] Tom White 著.曾大聃，周傲英译.Hadoop 权威指南[M]. 北京：清华大学出版社，2010.

[74] C. Fay, D. Jeffrey, G. Sanjay, H. Wilson C, W. Deborah A, B. Michael, C. Tushar,F. Andrew. "Bigtable: A Distributed Storage System for Structured Data". Research Google, 2006.

[75] 基于 HDFS 的云存储在高校信息资源整合中的应用，http://www.dzsc.com/data/html/2012-2-23/99979.html.

[76] http://www.cnad.com/html/Article/2013/0106/20130106175119592.shtml.

[77] Mary Meeker, 2012 Internet Trends.